Finite Wordlength Controller Realizations using the Specialized Implicit Form

Thibault Hilaire, Philippe Chevrel, James Whidborne

I R I S A
INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES

PUBLICATION
INTERNE
N° 1915

FINITE WORDLENGTH CONTROLLER REALIZATIONS USING THE SPECIALIZED IMPLICIT FORM

THIBAULT HILAIRE, PHILIPPE CHEVREL, JAMES F. WHIDBORNE

ISSN 1166-8687

IRISA
CAMPUS UNIVERSITAIRE DE BEAULIEU - 35042 RENNES CEDEX - FRANCE

INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES
Campus de Beaulieu – 35042 Rennes Cedex – France
Tél. : (33) 02 99 84 71 00 – Fax : (33) 02 99 84 71 71
http://www.irisa.fr

Finite Wordlength Controller Realizations using the Specialized Implicit Form

Thibault HILAIRE[*] , Philippe CHEVREL[**] , James F. WHIDBORNE[***]

Systèmes numériques
Projet CAIRN

Publication interne n ° 1915 — August 2008 — 39 pages

Abstract: A specialized implicit state-space representation is introduced to deal with finite wordlength effects in controller implementations. This specialized implicit form provides a macroscopic description of the algorithm to be implemented. So, it constitutes a unifying framework, allowing to encompass various implementation forms, such as the δ-operator, the ρDirect Form II transposed, observer-based and many other realizations considered usually separately in the literature. Different measures quantifying the finite wordlength effects on the overall closed loop behaviour, are defined in this new context. They concern both stability and performance. The gap with the infinite precision case is evaluated classically through the coefficient sensitivity and roundoff noise analysis. The problem of determining a realization with minimum finite wordlength effects can subsequently be solved using appropriate numerical methods. The approach is illustrated with two examples.

Key-words: Digital Control, Finite Wordlength Effects, Digital Controller Implementation, Optimal Realization

(Résumé : tsvp)

[*] CAIRN Project, INRIA/IRISA - France
[**] IRCCyN, UMR CNRS 6597, 1 rue de la Noë, 44321 Nantes, France
[***] Department of Aerospace Science, Cranfield University, UK

Centre National de la Recherche Scientifique
(UMR 6074) Université de Rennes 1 – Insa de Rennes
Institut National de Recherche en Informatique
et en Automatique – unité de recherche de Rennes

Modélisation des régulateurs avec la forme implicite spécialisée, en vue d'une implantation à précision finie.

Résumé : Une forme d'état implicite spécialisée est présentée pour étudier les effets de l'implantation en précision finie des régulateurs. Cette forme permet une description macroscopique des algorithmes à implanter. Elle constitue un canevas unificateur permettant de décrire les différentes structures utilisées pour l'implantation, telles que les réalisations avec l'opérateur δ, la forme directe II en ρ, la forme d'état-observateur et bien d'autres formes qui sont d'habitude traitées séparément dans la littérature. Différentes mesures quantifiant les effets de l'implantation sur le comportement en boucle fermée sont définis dans ce contexte. Elles concernent aussi bien la stabilité que la performance. L'écart entre la réalisation à précision infinie et la réalisation à précision finie est évaluée selon la mesure de sensibilité des coefficients et la mesure du bruit de quantification. Le problème consistant à trouver une réalisation dont l'implantation amène un minimum de dégradation peut alors est résolut numériquement. Cette approche est illustrée avec deux exemples.

Mots clés : Commande numérique, Précision finie, Implantation numérique de régulateur, réalisations optimales

1 Introduction

When implemented in digital computing devices, controllers are subjected to numerical degradations due to the rounding and quantization that occurs on the variables and constants used to define the controller. There are two main effects of this finite-precision (often known as the *Finite Word Length (FWL) effects*):

- the *roundoff noise* is the addition of noise into the system resulting from the rounding of variables before and after each arithmetic operation;
- the *parametric errors* are the quantization of the controller coefficients / parameters. They degrade the performance and/or stability of the controller.

For most low-order controllers, the FWL effects are minor, but for higher-order controllers, particularly when fast sampling is used, the FWL effects can become significant. For example, the stability of the system can be compromised even by a small quantization of the coefficients [30].

However, it is well-known that the FWL effects are dependent upon the controller realization. Hence many papers deal with the problem of finding a realization that minimize the FWL effects in some sense [see, for example, 5, 18, 29, and references therein]. It is also well-known that the FWL effects are dependent on the operator used. The δ-operator, for example, generally has much better numerical properties than the usual delay operator, q^{-1}, for control systems with fast sampling [6].

The problem of addressing the optimal realization for minimal FWL effects is usually addressed in the state space [e.g. 27, 5, 29]. Briefly, if the controller is

$$K(\sigma) = C(\sigma I - A)^{-1}B + D \tag{1}$$

where σ is usually the transform of the operator chosen (e.g. δ or q-operator), the problem is to search over the set

$$\left\{CT(\sigma I - T^{-1}AT)^{-1}TB + D : T \text{ a non-singular matrix}\right\}$$

to find a matrix T and corresponding controller realization with a small FWL effects. The limitations of this approach are that:

- there are many realizations that cannot be expressed in such a standard state space form;
- the search is restricted to a single operator.

The δ-operator is more complex to implement than the q-operator, so in some circumstances, it may be better to have a mix of operators. These limitations may be overcome by using the *Specialized Implicit Form* (SIF) [10] for the controller. The SIF allows a formal and faithful macroscopic description of the numerical algorithm used to implement the controller.

In order to determine the optimal realization, some measures of the roundoff noise and the closed-loop coefficient sensitivity are required. A fair number of these have been proposed over the years. The roundoff noise is generally measured by the output noise variance [for example, 24, 14, 5]. Measures of the input-output performance deterioration have been proposed by [5]. Stability can be assessed using a probabilistic measure [4], a measure based on a small-gain theorem [30] or closed-loop pole sensitivity measures [21, 29, 33, 19]. Ideally, the chosen measures should be computationally tractable but reasonably representative of the actual perturbations that occur in implementation.

The SIF was originally proposed in [10]. In [12] the FWL filter problem (the open-loop case) is considered. In this paper, some of the results of [12, 13] are extended to the FWL controller problem, that is the closed-loop case. A closed-loop input-output sensitivity measure which extends that of [5] and a Pole Sensitivity Stability Related Measure (PSSM) are proposed along with a closed-loop roundoff noise gain measure. All are suitable for use with the specialized implicit form and are similar to those proposed for the FWL filter realization problem [12]. Note that some preliminary results on FWL controller with the SIF appeared in [9].

The paper is organized as follows. In the next section, the SIF is recalled, and a number of definitions given. The recently proposed ρDFIIt realization [23] is shown to be a particular case of the SIF. In Section 3, the concept of equivalent classes (potentially structured) of realizations is introduced and illustrated by an example. Section 4 details, in a closed-loop context, the two sensitivity measures and the roundoff noise measure. In Section 5, an optimal design problem is introduced and it is illustrated with some examples in Section 6.

2 The Specialized Implicit Form

Many controller/filter forms, such as lattice filters and δ-operator controllers, make use of intermediate variables, and hence cannot be expressed in the traditional state-space form. The SIF has been proposed in order to model a much wider class of discrete-time linear time-invariant controller implementations than the classical state-space form.

The model takes the form of an implicit state-space realization [1] specialized according to

$$\begin{pmatrix} J & 0 & 0 \\ -K & I_n & 0 \\ -L & 0 & I_p \end{pmatrix} \begin{pmatrix} T(k+1) \\ X(k+1) \\ Y(k) \end{pmatrix} = \begin{pmatrix} 0 & M & N \\ 0 & P & Q \\ 0 & R & S \end{pmatrix} \begin{pmatrix} T(k) \\ X(k) \\ U(k) \end{pmatrix} \tag{2}$$

where $J \in \mathbb{R}^{l\times l}$, $K \in \mathbb{R}^{n\times l}$, $L \in \mathbb{R}^{p\times l}$, $M \in \mathbb{R}^{l\times n}$, $N \in \mathbb{R}^{l\times m}$, $P \in \mathbb{R}^{n\times n}$, $Q \in \mathbb{R}^{n\times m}$, $R \in \mathbb{R}^{p\times n}$, $S \in \mathbb{R}^{p\times m}$, $T(k) \in \mathbb{R}^l$, $X(k) \in \mathbb{R}^n$, $U(k) \in \mathbb{R}^m$ and $Y(k) \in \mathbb{R}^p$, and the matrix J is lower triangular with 1's on the main diagonal. Note $X(k+1)$ is the state-vector and is stored from one step to the next the vector, whilst T plays a particular role as $T(k+1)$ is independent of $T(k)$ (it is here defined as the vector of intermediary variables). The particular structure of J allows to express how the computations are decomposed with intermediates results that could be reused.

It is implicitly assumed throughout the paper that the computations associated with the realization (2) are executed in row order, giving the following algorithm:

$$\begin{aligned} &\text{[i]} \quad J.T(k+1) \leftarrow M.X(k) + N.U(k) \\ &\text{[ii]} \quad X(k+1) \leftarrow K.T(k+1) + P.X(k) + Q.U(k) \\ &\text{[iii]} \quad Y(k) \leftarrow L.T(k+1) + R.X(k) + S.U(k) \end{aligned} \tag{3}$$

Note that in practice, steps [ii] and [iii] could be exchanged to reduce the computational delay. Also note that because the computations are executed in row order and J is lower triangular with 1's on the main diagonal, there is no need to compute J^{-1}.

Equation (2) is equivalent in infinite precision to the classical state-space form

$$\left(\begin{array}{c} T(k+1) \\ X(k+1) \\ \hline Y(k) \end{array}\right) = \left(\begin{array}{cc|c} 0 & J^{-1}M & J^{-1}N \\ 0 & A_Z & B_Z \\ \hline 0 & C_Z & D_Z \end{array}\right) \left(\begin{array}{c} T(k) \\ X(k) \\ \hline U(k) \end{array}\right) \tag{4}$$

with $A_Z \in \mathbb{R}^{n\times n}$, $B_Z \in \mathbb{R}^{n\times m}$, $C_Z \in \mathbb{R}^{p\times n}$ and $D_Z \in \mathbb{R}^{p\times m}$ where

$$A_Z = KJ^{-1}M + P, \qquad B_Z = KJ^{-1}N + Q, \tag{5}$$

$$C_Z = LJ^{-1}M + R, \qquad D_Z = LJ^{-1}N + S. \tag{6}$$

Note that (4) corresponds to a different parametrization than (2) (the finite-precision implementation of (4) will cause different numerical deterioration to that of (2)). The associated system transfer function is given by

$$H : z \mapsto C_Z(zI_n - A_Z)^{-1}B_Z + D_Z. \tag{7}$$

A complete framework for the description of all digital controller implementations can be developed by using the following definitions. For further details, see [12].

Definition 1 *A **realization** $\mathcal{R}$ of a transfer matrix H is entirely defined by the data Z, l, m, n and p. $Z \in \mathbb{R}^{(l+n+p)\times(l+n+m)}$ is partitioned according to*

$$Z \triangleq \begin{pmatrix} -J & M & N \\ K & P & Q \\ L & R & S \end{pmatrix} \tag{8}$$

and l, m, n and p are the matrix dimensions given previously. The notation used will be $\mathcal{R} := (Z, l, m, n, p)$.

The notation Z is introduced to make the further developments more compact (see (44), (61), etc.).

Definition 2 *$\mathcal{R}_H$ denotes the set of realizations described by* (2) *equivalent to the transfer function H, that is to say with the same input-output relationship. These realizations are said to be Input-Output equivalent(IO-equivalent) and Input-Output equivalent to the transfer function H.*

In order to encompass realizations with some special structure (q or δ state-space, direct forms, cascades, lattice, etc.), a subset of realizations sharing the same structure is defined.

Definition 3 *A **structuration** $\mathbb{S}$ is a set of structured realizations. That is realizations that share a common structure with some coefficients and/or some dimensions having been fixed* a priori.

Some examples of structurations are given in the next sub-section.

Definition 4 *$\mathcal{R}_H^{\mathbb{S}}$ is the set of equivalent structured realizations. Realizations from $\mathcal{R}_H^{\mathbb{S}}$ are structured according to $\mathbb{S}$ and are IO-equivalent to H:*

$$\mathcal{R}_H^{\mathbb{S}} \triangleq \mathcal{R}_H \cap \mathbb{S}. \tag{9}$$

2.1 Some examples

2.1.1 δ-realizations

Consider the δ-state-space form

$$\begin{cases} \delta[X(k)] = A_\delta X(k) + B_\delta U(k) \\ Y(k) = C_\delta X(k) + D_\delta U(k) \end{cases} \tag{10}$$

with $\delta = \frac{q-1}{\Delta}$, $\Delta \in \mathbb{R}_{+*}$ and q is the shift operator [5].

This realization should be implemented with the following algorithm

$$\begin{array}{ll} \text{[i]} & T \leftarrow A_\delta.X(k) + B_\delta.U(k) \\ \text{[ii]} & X(k+1) \leftarrow X(k) + \Delta.T \\ \text{[iii]} & Y(k) \leftarrow C_\delta.X(k) + D_\delta.U(k) \end{array} \tag{11}$$

where T is an intermediate variable. This could be modelled with the specialized implicit form as

$$\begin{pmatrix} I_n & 0 & 0 \\ -\Delta I_n & I_n & 0 \\ 0 & 0 & I_p \end{pmatrix} \begin{pmatrix} T(k+1) \\ X(k+1) \\ Y(k) \end{pmatrix} = \begin{pmatrix} 0 & A_\delta & B_\delta \\ 0 & I_n & 0 \\ 0 & C_\delta & D_\delta \end{pmatrix} \begin{pmatrix} T(k) \\ X(k) \\ U(k) \end{pmatrix} \tag{12}$$

So, the δ-structuration, $\mathbb{S}_\delta$, is formally defined by

$$\mathbb{S}_\delta = \left\{ \begin{array}{l} \mathcal{R} := (I_n, \Delta I_n, 0, A_\delta, B_\delta, I_n, 0, C_\delta, D_\delta) \\ \quad \forall m \in \mathbb{N}, n \in \mathbb{N}, p \in \mathbb{N} \\ \quad \forall \Delta \in \mathbb{R}^+, A_\delta \in \mathbb{R}^{n\times n}, B_\delta \in \mathbb{R}^{n\times m} \\ \quad \forall C_\delta \in \mathbb{R}^{p\times n}, D_\delta \in \mathbb{R}^{p\times m} \end{array} \right\} \tag{13}$$

2.1.2 Cascade decomposition

The cascade form is a common realization for filter/controller implementations. It generally has good FWL properties compared to the direct forms and requires less operations than fully parametrized state-space realizations. The system is decomposed into a number of lower order (usually first and second-order) subsystems connected in series.

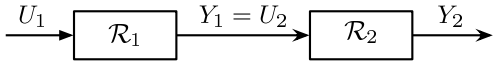

Figure 1: Cascade form

Let us consider two realizations $\mathcal{R}_1$ and $\mathcal{R}_2$ connected in series as shown in Figure 1. Assuming $\mathcal{R}_1$ and $\mathcal{R}_2$ to be defined by SIF matrices $(J_1, K_1, L_1, M_1, N_1, P_1, Q_1, R_1, S_1)$ and $(J_2, K_2, L_2, M_2, N_2, P_2, Q_2, R_2, S_2)$, and cascading them leads to the realization $\mathcal{R} := (Z, m_1, p_1 + l_1 + l_2, n_1 + n_2, p_2)$ with

$$Z = \left(\begin{array}{ccc:ccc} -J_1 & 0 & 0 & M_1 & 0 & N_1 \\ L_1 & -I & 0 & R_1 & 0 & S_1 \\ 0 & N_2 & -J_2 & 0 & M_2 & 0 \\ \hdashline K_1 & 0 & 0 & P_1 & 0 & Q_1 \\ 0 & Q_2 & K_2 & 0 & P_2 & 0 \\ \hdashline 0 & S_2 & L_2 & 0 & R_2 & 0 \end{array}\right) \tag{14}$$

from which definition of the corresponding structuration $\mathcal{S}$ immediately follows. The outputs of $\mathcal{R}_1$ are computed in the intermediate variable, and then used as the inputs of $\mathcal{R}_2$.

The main point is that this construction can represent cascade systems without changing the parametrization.

Remark 1 The cascade structuration can be applied to realizations that are structured differently (q and δ-state-space realizations for example) and easily extended to multiple cascaded systems.

2.1.3 ρ Transposed Direct-form II

Li and Hao [23, 7, 22] have presented a new sparse structure called ρDFIIt. This is a generalization of the transposed direct-form II structure with the conventional shift and the δ-operator and is similar to that of [25]. It is a sparse realization (with $3n + 1$ parameters when n is the order of the controller), leading so to an economic (few computations) implementation that could be very numerically efficient. As we will see later, this realization has n extra degrees of freedom that can be used to find an *optimal* realization within its particular structuration.

Let us define

$$\rho_i : z \mapsto \frac{z-\gamma_i}{\Delta_i}, \qquad 1 \leqslant i \leqslant n \tag{15}$$

and

$$\varrho_i : z \mapsto \prod_{j=1}^{i} \rho_j(z), \qquad 1 \leqslant i \leqslant n \tag{16}$$

where $(\gamma_i)_{1\leqslant i\leqslant n}$ and $(\Delta_i > 0)_{1\leqslant i\leqslant n}$ are two sets of constants. Let $(a_i)_{1\leqslant i\leqslant n}$ and $(b_i)_{0\leqslant i\leqslant n}$ be the coefficient sets of the transfer function, using the shift operator:

$$H : z \mapsto \frac{b_0 + b_1 z^{-1} + \ldots + b_{n-1} z^{-n+1} + b_n z^{-n}}{1 + a_1 z^{-1} + \ldots + a_{n-1} z^{-n+1} + a_n z^{-n}} \tag{17}$$

Therefore, H can be **reparametrized** with $(\alpha_i)_{1\leqslant i\leqslant n}$ and $(\beta_i)_{0\leqslant i\leqslant n}$ as follows:

$$H(z) = \frac{\beta_0 + \beta_1 \varrho_1^{-1}(z) + \ldots + \beta_{n-1}\varrho_{n-1}^{-1}(z) + \beta_n \varrho_n^{-1}(z)}{1 + \alpha_1 \varrho_1^{-1}(z) + \ldots + \alpha_{n-1}\varrho_{-1}^{-n+1}(z) + \alpha_n \varrho_n^{-1}(z)} \tag{18}$$

Denoting

$$V_a \triangleq \begin{pmatrix} 1 \\ a_1 \\ \vdots \\ a_n \end{pmatrix}, V_b \triangleq \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{pmatrix}, V_\alpha \triangleq \begin{pmatrix} 1 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}, V_\beta \triangleq \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} \tag{19}$$

the parameters $(a_i)_{1\leqslant i\leqslant n}$, $(b_i)_{0\leqslant i\leqslant n}$, $(\alpha_i)_{1\leqslant i\leqslant n}$ and $(\beta_i)_{0\leqslant i\leqslant n}$ are related [7] according to:

$$\begin{cases} V_a = \kappa \Omega V_\alpha \\ V_b = \kappa \Omega V_\beta \end{cases} \tag{20}$$

where $\kappa \triangleq \prod_{i=1}^{n} \Delta_i$ and $\Omega \in \mathbb{R}^{n+1\times n+1}$ is a lower triangular matrix whose ith column is determined by the coefficients of the z-polynomial $\prod_{j=i}^{n} \rho_j(z)$ for $1 \leqslant i \leqslant n$ and $\Omega_{n+1,n+1} = 1$.

Equation (18) can be, for example, implemented with a transposed direct form II (see Figure 2), and each operator ρ_i^{-1} can be implemented as shown in Figure 3 (each ϱ_k^{-1} is obtained by cascading the $\left(\rho_i^{-1}\right)_{1\leqslant i\leqslant k}$). Clearly, when $\gamma_i = 0$, $\Delta_i = 1 \quad (1 \leqslant i \leqslant n)$, Figure 2 is the conventional transposed direct form II. When $\gamma_i = 1$, $\Delta_i = \Delta \quad (1 \leqslant i \leqslant n)$, one gets the δ transposed direct form II. This form was first proposed as an unification for the shift-direct form II transposed and the δ-direct form II transposed. It is now used to exploit the n extra degrees of freedom given by the choice of the parameters $(\gamma_i)_{1\leqslant i\leqslant n}$.

The corresponding algorithm is:

$$\begin{aligned} &\text{[i]} \quad Y(k) \leftarrow \beta_0 U(k) + W_1(k) \\ &\text{[ii]} \quad W_i(k) \leftarrow \rho_i^{-1}\left[\beta_i U(k) - \alpha_i Y(k) + W_{i+1}(k)\right] \\ &\text{[iii]} \quad W_n(k) \leftarrow \rho_n^{-1}\left[\beta_n U(k) - \alpha_n Y(k)\right] \end{aligned} \tag{21}$$

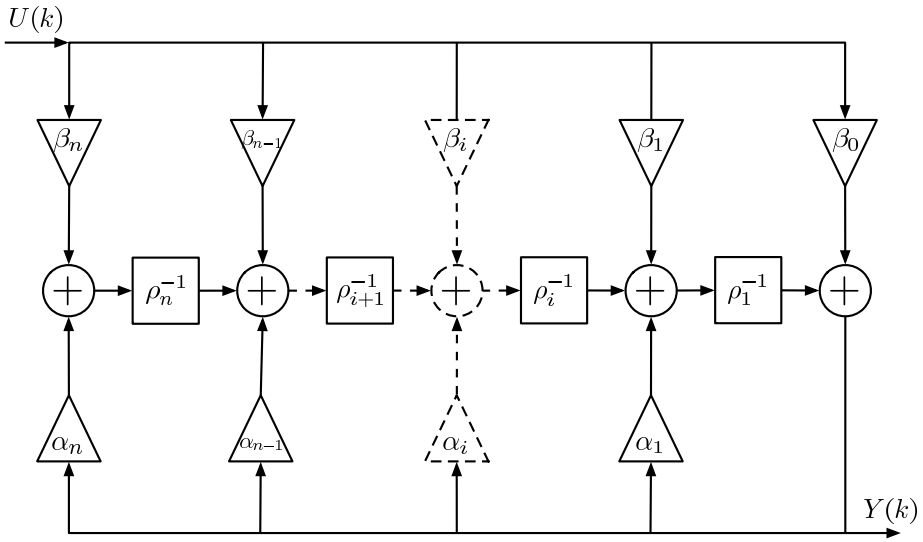

Figure 2: Generalized ρ Direct Form II

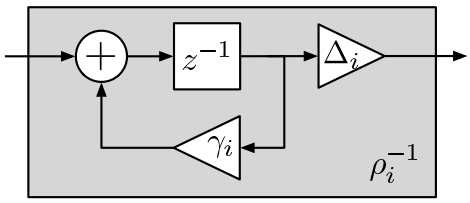

Figure 3: Realization of operator ρ_i^{-1}

By introducing the intermediate variables needed to realize the ρ_i^{-1} operator (according to $\rho_i^{-1} = \frac{1}{q^{-1}-\gamma_i}\Delta_i$, with the multiplication by Δ_i done last, see Figure 3), equations (22) to (24) become

$$T = \begin{pmatrix} \Delta_1 & & & \\ & \Delta_2 & & \\ & & \ddots & \\ & & & \Delta_n \end{pmatrix} X(k) + \begin{pmatrix} \beta_0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} U(k) \tag{22}$$

$$X(k+1) = \begin{pmatrix} -\alpha_1 & 1 & & \\ -\alpha_2 & 0 & \ddots & \\ \vdots & & \ddots & 1 \\ -\alpha_n & & & 0 \end{pmatrix} T + \begin{pmatrix} \gamma_1 & & & \\ & \gamma_2 & & \\ & & \ddots & \\ & & & \gamma_n \end{pmatrix} X(n) + \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} U(k) \tag{23}$$

$$Y(k) = \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix} T \tag{24}$$

Within the SIF Framework, the ρDFIIt form is described by:

$$Z = \left(\begin{array}{cccc:ccccc} -1 & & & & \Delta_1 & & & & \beta_0 \\ & \ddots & & & & \Delta_2 & & & 0 \\ & & \ddots & & & & \ddots & & \vdots \\ & & & -1 & & & & \Delta_n & 0 \\ \hdashline -\alpha_1 & 1 & & & \gamma_1 & & & & \beta_1 \\ -\alpha_2 & 0 & \ddots & & & \gamma_2 & & & \beta_2 \\ \vdots & & \ddots & 1 & & & \ddots & & \vdots \\ -\alpha_n & & & 0 & & & & \gamma_n & \beta_n \\ \hdashline 1 & 0 & \dots & 0 & 0 & \dots & \dots & 0 & 0 \end{array}\right) \tag{25}$$

Remark 2 Thanks to the SIF, there is no need to use another operator unlike the shift operator.

A number of other examples of structurations are given in [8]. They illustrate the generality of the SIF framework.

3 Equivalent classes

In order to exploit the potential offered by the specialized implicit form in improving implementations, it is necessary to characterize further the sets of equivalent system realizations. We firstly note that non-minimal realizations may provide better implementations (the δ-form can be seen as a non-minimal realization when written in the implicit state-space form

– with the shift operator). Hence the notion of equivalence needs to be extended by considering that the system state dimension does not have to be invariant. The *Inclusion Principle*, introduced by Šiljak and Ikeda [see 16, 28] in the context of decentralized control, is useful here as it allows the formalization of the *equivalence* and *inclusion* relations between two system realizations.

Definition 5 *Consider two systems Σ and $\tilde{\Sigma}$, with state dimension n and $\tilde{n} \geqslant n$ respectively, described in the classical state-space form by the matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, $\tilde{A} \in \mathbb{R}^{\tilde{n}\times\tilde{n}}$, $\tilde{B} \in \mathbb{R}^{\tilde{n}\times m}$ and $\tilde{C} \in \mathbb{R}^{p\times\tilde{n}}$. The system Σ is said to be* ***included*** *in the system $\tilde{\Sigma}$ (denoted by $\Sigma \subset \tilde{\Sigma}$) if there exists $(\mathcal{U}, \mathcal{V}) \in \mathbb{R}^{n\times\tilde{n}} \times \mathbb{R}^{\tilde{n}\times n}$ such that $\mathcal{U}\mathcal{V} = I_n$ and, for any initial state $X(0) = X_0$ of Σ and any input $(U(k))_{k\geqslant 0}$, the choice of the initial state $\tilde{X}(0) = \mathcal{V}X_0$ of $\tilde{\Sigma}$ implies*

$$\begin{cases} X(k) &= \mathcal{U}\tilde{X}(k) \\ Y(k) &= \tilde{Y}(k) \end{cases} \qquad \forall k \geqslant 0. \tag{26}$$

Remark 3 Equation (26) implies that system $\tilde{\Sigma}$ contains all the information to get the trajectory of Σ.

The principle is extended here to the specialized implicit form in order to characterize equivalence classes. An equivalence class is defined by a certain minimal realization and all the realizations that include this realization. They can be constructed using the following proposition:

Proposition 1 *Consider a realization $\mathcal{R} := (J, K, L, M, N,\ P, Q, R, S)$ with dimensions l, m, n, p. A realization $\tilde{\mathcal{R}}$ that includes $\mathcal{R}$ can be constructed as follows:*

- *Choose $\tilde{n}$ and $\tilde{l}$ such that $\tilde{n} + \tilde{l} \geqslant n + l$*
- *Choose $(\mathcal{U}, \mathcal{V}) \in \mathbb{R}^{n\times\tilde{n}} \times \mathbb{R}^{\tilde{n}\times n}$ such that $\mathcal{U}\mathcal{V} = I_n$, $(\mathcal{W}, \mathcal{T}) \in \mathbb{R}^{l\times\tilde{l}} \times \mathbb{R}^{\tilde{l}\times l}$ such that $\mathcal{W}\mathcal{T} = I_l$ and $(\mathcal{X}, \mathcal{Y}) \in \mathbb{R}^{l\times\tilde{l}} \times \mathbb{R}^{\tilde{l}\times l}$ such that $\mathcal{X}\mathcal{Y} = I_l$.*
- *Choose complementary matrices*[1] *$\mathcal{M}_{\tilde{J}^{-1}} \in \mathbb{R}^{\tilde{l}\times\tilde{l}}$, $\mathcal{M}_{\tilde{K}} \in \mathbb{R}^{\tilde{n}\times\tilde{l}}$, $\mathcal{M}_{\tilde{L}} \in \mathbb{R}^{p\times\tilde{l}}$, $\mathcal{M}_{\tilde{M}} \in \mathbb{R}^{\tilde{l}\times\tilde{n}}$, $\mathcal{M}_{\tilde{N}} \in \mathbb{R}^{\tilde{l}\times m}$, $\mathcal{M}_{\tilde{P}} \in \mathbb{R}^{\tilde{n}\times\tilde{n}}$, $\mathcal{M}_{\tilde{Q}} \in \mathbb{R}^{\tilde{n}\times m}$, $\mathcal{M}_{\tilde{R}} \in \mathbb{R}^{p\times\tilde{n}}$ and $\mathcal{M}_{\tilde{S}} \in \mathbb{R}^{p\times m}$ such that, if we denote $\tilde{J}^{-1} = \mathcal{T}J^{-1}\mathcal{X} + \mathcal{M}_{\tilde{J}^{-1}}$, $\tilde{K} = \mathcal{V}K\mathcal{W} + \mathcal{M}_{\tilde{K}}$, $\tilde{L} = L\mathcal{W} + \mathcal{M}_{\tilde{L}}$, $\tilde{M} = \mathcal{Y}M\mathcal{U} + \mathcal{M}_{\tilde{M}}$, $\tilde{N} = \mathcal{Y}N + \mathcal{M}_{\tilde{N}}$, $\tilde{P} = \mathcal{V}P\mathcal{U} + \mathcal{M}_{\tilde{P}}$, $\tilde{Q} = \mathcal{V}Q + \mathcal{M}_{\tilde{Q}}$, $\tilde{R} = R\mathcal{U} + \mathcal{M}_{\tilde{R}}$, $\tilde{S} = S + \mathcal{M}_{\tilde{S}}$ and*

$$\begin{pmatrix} \mathcal{M}_{\tilde{A}} & \mathcal{M}_{\tilde{B}} \\ \mathcal{M}_{\tilde{C}} & \mathcal{M}_{\tilde{D}} \end{pmatrix} = \begin{pmatrix} K \\ L \end{pmatrix} J^{-1} \begin{pmatrix} M & N \end{pmatrix} + \begin{pmatrix} P & Q \\ R & S \end{pmatrix} \begin{pmatrix} \tilde{K} \\ \tilde{L} \end{pmatrix} \tilde{J}^{-1} \begin{pmatrix} \tilde{M} & \tilde{N} \end{pmatrix} + \begin{pmatrix} \tilde{P} & \tilde{Q} \\ \tilde{R} & \tilde{S} \end{pmatrix} \tag{27}$$

[1] These matrices are called *complementary matrices*. $\mathcal{M}_{\tilde{X}}$ is complementary in that it fills the gap between $\tilde{X}$ and the similarity on X: $\tilde{X} = \mathcal{T}_1 X \mathcal{T}_2 + \mathcal{M}_{\tilde{X}}$.

then $\mathcal{U}\left(\mathcal{M}_{\tilde{A}}\right)^i\mathcal{V} = 0\ \forall i \geqslant 1$, $\mathcal{U}\left(\mathcal{M}_{\tilde{A}}\right)^i\mathcal{M}_{\tilde{B}} = 0\ \forall i \geqslant 0$, $\mathcal{M}_{\tilde{C}}\left(\mathcal{M}_{\tilde{A}}\right)^i\mathcal{V} = 0\ \forall i \geqslant 0$, $\mathcal{M}_{\tilde{C}}\left(\mathcal{M}_{\tilde{A}}\right)^i\mathcal{M}_{\tilde{B}} = 0\ \forall i \geqslant 0$ *and* $\mathcal{M}_{\tilde{D}} = 0$ *are satisfied.*

If so, the realization $\tilde{\mathcal{R}} := (\tilde{J}, \tilde{K}, \tilde{L}, \tilde{M}, \tilde{N}, \tilde{P}, \tilde{Q}, \tilde{R}, \tilde{S})$ *includes the realization* $\mathcal{R}$.
Proof:
The proof can be derived directly from the characterization of the Inclusion Principle [15, 16, 2]. The details are omitted here but can be found in [8]. ■

Although this extension gives the formal description of equivalent classes, it may be of practical interest to consider realizations of the same dimensions ($\tilde{l} = l$ and $\tilde{n} = n$), where transformations from one realization to another is only a similarity transformation.

Proposition 2 *Consider a realization* $\mathcal{R} := (Z, l, m, n, p)$. *All the realizations* $\tilde{\mathcal{R}} := (\tilde{Z}, l, m, n, p)$ *with*

$$\tilde{Z} = \begin{pmatrix} \mathcal{Y} & & \\ & \mathcal{U}^{-1} & \\ & & I_p \end{pmatrix} Z \begin{pmatrix} \mathcal{W} & & \\ & \mathcal{U} & \\ & & I_m \end{pmatrix} \tag{28}$$

and $\mathcal{U}$, $\mathcal{W}$, $\mathcal{Y}$ *are non-singular matrices, are equivalent to* $\mathcal{R}$, *and share the same complexity (i.e. generically the same amount of computation).*

It is also possible to just consider a subset of similarity transformations that preserve a particular structure, say cascade or delta. For example, if an initial δ-structured realization $\mathcal{R} := (Z_0, n, m, n, p)$ is given, the subset of equivalent δ-structured realization is defined by

$$\mathcal{R}_H^{\mathcal{S}_\delta} = \left\{ \begin{array}{l} \mathcal{R} := (Z, n, m, n, p) \backslash \\ \quad Z = \begin{pmatrix} \mathcal{U}^{-1} & & \\ & \mathcal{U}^{-1} & \\ & & I_p \end{pmatrix} Z_0 \begin{pmatrix} \mathcal{U} & & \\ & \mathcal{U} & \\ & & I_m \end{pmatrix} \\ \quad \forall \mathcal{U} \in \mathbb{R}^{n\times n} \text{ non-singular} \end{array} \right\} \tag{29}$$

This compact algebraic characterization of equivalent classes is particularly efficient when used to search for an optimal structured realization (see Section 5).

4 Closed-loop measures

The quantization of the coefficients and the roundoff noise may have a negative impact on the closed-loop system behaviour. Three measures that may be used to evaluate this impact are described in this section.

4.1 Problem statement

Consider the plant $\mathcal{P}$ together with controller $\mathcal{C}$ according to the standard form shown in Figure 4, where $W(k) \in \mathbb{R}^{p_1}$ is the exogenous input, $Y(k) \in \mathbb{R}^{p_2}$ the control input, $Z(k) \in \mathbb{R}^{m_1}$ the controlled output and $U(k) \in \mathbb{R}^{m_2}$ the measured output.

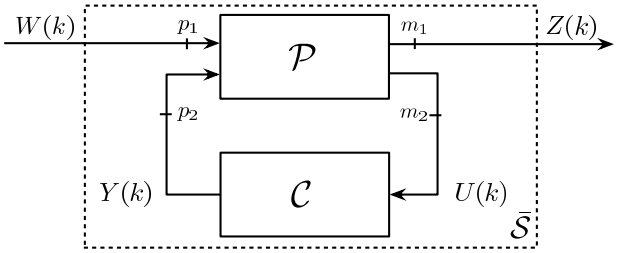

Figure 4: Closed-loop control system

The controller is defined as $\mathcal{C} := (Z, l, m_2, n, p_2)$ and the plant $\mathcal{P}$ as

$$\mathcal{P} := \left(\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & 0 \end{array}\right) \tag{30}$$

where $A \in \mathbb{R}^{n_\mathcal{P} \times n_\mathcal{P}}$, $B_1 \in \mathbb{R}^{n_\mathcal{P} \times p_1}$, $B_2 \in \mathbb{R}^{n_\mathcal{P} \times p_2}$, $C_1 \in \mathbb{R}^{m_1 \times n_\mathcal{P}}$, $C_2 \in \mathbb{R}^{m_2 \times n_\mathcal{P}}$, $D_{11} \in \mathbb{R}^{m_1 \times p_1}$, $D_{12} \in \mathbb{R}^{m_1 \times p_2}$, $D_{21} \in \mathbb{R}^{m_2 \times p_1}$ and $D_{22} \in \mathbb{R}^{m_2 \times p_2}$ is assumed to be zero only to simplify the mathematical expressions.

Note that open loop results (filter modelling) may be obtained as a particular case, with:

$$\mathcal{P} := \left(\begin{array}{c|cc} & & \\ \hline & 0 & I \\ & I & 0 \end{array}\right). \tag{31}$$

The closed-loop system $\bar{\mathcal{S}}$ is then given by

$$\bar{\mathcal{S}} = F_l(\mathcal{P}, \mathcal{C}) := \left(\begin{array}{c|c} \bar{A} & \bar{B} \\ \hline \bar{C} & \bar{D} \end{array}\right) \tag{32}$$

where $F_l(\cdot,\cdot)$ is the well-known lower linear fractional transform [35] and where $\bar{A} \in \mathbb{R}^{n_\mathcal{P}+n \times n_\mathcal{P}+n}$, $\bar{B} \in \mathbb{R}^{n_\mathcal{P}+n \times p_1}$, $\bar{C} \in \mathbb{R}^{m_1 \times n_\mathcal{P}+n}$ and $\bar{D} \in \mathbb{R}^{m_1 \times p_1}$ are such that

$$\bar{A} = \begin{pmatrix} A + B_2 D_Z C_2 & B_2 C_Z \\ B_Z C_2 & A_Z \end{pmatrix}, \qquad \bar{B} = \begin{pmatrix} B_1 + B_2 D_Z D_{21} \\ B_Z D_{21} \end{pmatrix}, \tag{33}$$

$$\bar{C} = \begin{pmatrix} C_1 + D_{12} D_Z C_2 & D_{12} C_Z \end{pmatrix}, \qquad \bar{D} = D_{11} + D_{12} D_Z D_{21}. \tag{34}$$

The closed-loop transfer function is

$$\bar{H} : z \mapsto \bar{C} \left(zI - \bar{A}\right)^{-1} \bar{B} + \bar{D}. \tag{35}$$

4.2 Input-output sensitivity

In order to evaluate how much the quantization of the controller's coefficients (due to FWL implementation) affects the closed-loop transfer function, the sensitivity $\frac{\partial \bar{H}}{\partial Z}$ can be used. Before that, the nature of the perturbation on each coefficient must made precise.

A coefficient's quantization depends both on its value and its representation. Firstly if the value of a coefficient is such that it will be quantized without error (like 0, ± 1 or a power of 2), then, that parameter makes no contribution to the overall coefficient sensitivity and is called a *trivial* parameter. Hence we introduce the weighting matrices W_Z associated with Z such that

$$(W_Z)_{i,j} \triangleq \begin{cases} 0 & \text{if } X_{i,j} \text{ is exactly implemented,} \\ 1 & \text{otherwise.} \end{cases} \tag{36}$$

For a fixed-point representation, Z is perturbed to $Z^\dagger = Z + W_Z \times \Delta$, where Δ represents the quantification error.

Remark 4 For floating-point representations, Z is perturbed to $Z^\dagger = Z + W_Z \times Z \times \Delta$ [34, 11]. The following measures can then be easily extended to the floating-point (and block-floating-point) case.

The closed-loop transfer function resulting from the quantization process is denoted by $\bar{H}^\dagger \triangleq \bar{H}\big|_{Z+W_Z\times\Delta}$. For the Single Input Single Output (SISO) case, the following is true $\forall z \in \mathbb{C}$

$$\bar{H}^\dagger(z) - \bar{H}(z) = \sum_{i,j} \Delta_{i,j} \left.\frac{\partial \bar{H}^\dagger(z)}{\partial \Delta}\right|_{\Delta=0} + o\left(\|\Delta\|^2_{\max}\right) \tag{37}$$

and

$$\left\|\bar{H}^\dagger - \bar{H}\right\|_2 \leq \|\Delta\|_{\max} \left\| \left.\frac{\partial \bar{H}^\dagger}{\partial \Delta}\right|_{\Delta=0} \right\|_2 + o\left(\|\Delta\|^2_{\max}\right) \tag{38}$$

where $\|\cdot\|_2$ denotes the H_2-norm.

It is easy to show that

$$\left.\frac{\partial \bar{H}^\dagger}{\partial \Delta}\right|_{\Delta=0} = \frac{\partial \bar{H}}{\partial Z} \times W_Z \tag{39}$$

From (38) and (39), we define an input-output sensitivity measure as follows:

Definition 6 *Consider a realization $\mathcal{C} := (Z, l, m_2, n, p_2)$. For the SISO case, the closed-loop transfer function sensitivity, with respect to all the non-trivial coefficients of $\mathcal{C}$, is defined by*

$$\bar{M}^W_{L_2} \triangleq \left\| \frac{\partial \bar{H}}{\partial Z} \times W_Z \right\|^2_2. \tag{40}$$

Remark 5 It is possible to include a frequency weighting to emphasize certain frequency range [5] to ensure that the closed-loop degradation is constrained over a given frequency range.

This measure can be extended to the Multiple Input Multiple Output (MIMO) case. It is also useful to consider the contribution of each coefficient to the overall sensitivity. The *closed-loop transfer function sensitivity matrix*, denoted by $\frac{\delta \bar{H}}{\delta Z}$, is the matrix of the H_2-norm of the input-output sensitivity of the transfer function $\bar{H}$ with respect to each coefficient $Z_{i,j}$. It is defined by

$$\left(\frac{\delta \bar{H}}{\delta Z}\right)_{i,j} \triangleq \left\| \frac{\partial \bar{H}}{\partial Z_{i,j}} \right\|_2 . \tag{41}$$

It can be used to obtain a *map* of the sensitivity with respect to each coefficient and help to choose a specific fixed-point format for each coefficient. From the properties of H_2-norms, we get

$$\left\| \frac{\delta \bar{H}}{\delta Z} \right\|_F = \left\| \frac{\partial \bar{H}}{\partial Z} \right\|_2 \tag{42}$$

where $\|\cdot\|_F$ is the Frobenius norm. Definition 6 can now be stated for the general case.

Definition 7 *The closed-loop input-output sensitivity measure is defined by*

$$\bar{M}_{L_2}^{W} \triangleq \left\| \frac{\delta \bar{H}}{\delta Z} \times W_Z \right\|_F^2 . \tag{43}$$

The input-output sensitivity $\frac{\partial \bar{H}}{\partial Z}$ can be evaluated by the following proposition.

Proposition 3

$$\frac{\partial \bar{H}}{\partial Z} = \bar{H}_1 \circledast \bar{H}_2 \tag{44}$$

where $\circledast$ *is the operator defined by*

$$A \circledast B \triangleq \mathrm{Vec}(A).\left[\mathrm{Vec}\left(B^{\top}\right)\right]^{\top}, \tag{45}$$

$\mathrm{Vec}(\cdot)$ *is the classical operator that vectorizes a matrix,* $\bar{H}_1$ *and* $\bar{H}_2$ *are defined by*

$$\bar{H}_1 : z \mapsto \bar{C}\left(zI - \bar{A}\right)^{-1} \bar{M}_1 + \bar{M}_2 \tag{46}$$

$$\bar{H}_2 : z \mapsto \bar{N}_1\left(zI - \bar{A}\right)^{-1} \bar{B} + \bar{N}_2 \tag{47}$$

and

$$\bar{M}_1 = \begin{pmatrix} B_2 L J^{-1} & 0 & B_2 \\ K J^{-1} & I_n & 0 \end{pmatrix}, \qquad \bar{N}_1 = \begin{pmatrix} J^{-1} N C_2 & J^{-1} M \\ 0 & I_n \\ C_2 & 0 \end{pmatrix}, \tag{48}$$

$$\bar{M}_2 = \begin{pmatrix} D_{12} L J^{-1} & 0 & D_{12} \end{pmatrix}, \qquad \bar{N}_2 = \begin{pmatrix} J^{-1} N D_{21} \\ 0 \\ D_{21} \end{pmatrix}. \tag{49}$$

The dimensions of $\bar{M}_1$, $\bar{M}_2$, $\bar{N}_1$ and $\bar{N}_2$ are respectively $(n+n_{\mathcal{P}})\times(l+n+p_2)$, $m_1\times(l+n+p_2)$, $(l+n+m_2)\times(n+n_{\mathcal{P}})$ and $(l+n+m_2)\times p_1$.
Proof:
The proof is based on the following lemma and can be found in [11, 8].

Lemma 1 *Let X be a matrix in $\mathbb{R}^{p\times l}$ while G and H are two transfer matrices independent of X with values in $\mathbb{C}^{m\times p}$ and $\mathbb{C}^{l\times n}$ respectively and that are independent of X. Then*

$$\frac{\partial(GXH)}{\partial X} = G \circledast H, \tag{50}$$

$$\frac{\partial(GX^{-1}H)}{\partial X} = (GX^{-1}) \circledast (X^{-1}H). \tag{51}$$

From (33), (5) and (6), it is possible to write

$$\bar{A} = \begin{pmatrix} A + B_2LJ^{-1}NC_2 & B_2C_Z \\ B_ZC_2 & A_Z \end{pmatrix} + \begin{pmatrix} B_2 \\ 0 \end{pmatrix} S \begin{pmatrix} C_2 & 0 \end{pmatrix} \tag{52}$$

and finally with Lemma 1

$$\frac{\partial \bar{H}}{\partial S} = \begin{pmatrix} B_2 \\ 0 \end{pmatrix} \circledast \begin{pmatrix} C_2 & 0 \end{pmatrix}. \tag{53}$$

The other derivatives $\frac{\partial \bar{H}}{\partial R}$, $\frac{\partial \bar{H}}{\partial Q}$, ... can be similarly obtained and then gathered using:

$$\frac{\partial}{\partial Z} = \begin{pmatrix} -\frac{\partial}{\partial J} & \frac{\partial}{\partial M} & \frac{\partial}{\partial N} \\ \frac{\partial}{\partial K} & \frac{\partial}{\partial P} & \frac{\partial}{\partial Q} \\ \frac{\partial}{\partial L} & \frac{\partial}{\partial R} & \frac{\partial}{\partial S} \end{pmatrix}. \tag{54}$$

■

Proposition 4 *The closed-loop transfer function sensitivity matrix $\frac{\delta \bar{H}}{\delta Z}$ can be computed as*

$$\left(\frac{\delta \bar{H}}{\delta Z}\right)_{i,j} = \left\| \bar{H}_1 E_{i,j} \bar{H}_2 \right\|_2 \tag{55}$$

with

$$\bar{H}_1 E_{i,j} \bar{H}_2 := \left(\begin{array}{cc|c} \bar{A} & 0 & \bar{B} \\ \bar{M}_1 E_{i,j} \bar{N}_1 & \bar{A} & \bar{M}_1 E_{i,j} \bar{N}_2 \\ \hline \bar{M}_2 E_{i,j} \bar{N}_1 & \bar{C} & \bar{M}_2 E_{i,j} \bar{N}_2 \end{array}\right) \tag{56}$$

and $E_{i,j}$ is the matrix of appropriate size with all elements being 0 except the (i,j)th element which is unity.
Proof:
The proof is quite straightforward, and comes from the definition of operator $\circledast$ in Proposition 3. ■

Remark 6 In the SISO case, the problem becomes simpler by noting that

$$\left(\frac{\delta \bar{H}}{\delta Z}\right)_{i,j} = \left\|(H_2 H_1)_{i,j}\right\|_2 \tag{57}$$

$$= \left\| \left(\begin{array}{cc|c} \bar{A} & 0 & \bar{B} \\ \bar{M}_1\bar{N}_1 & \bar{A} & \bar{M}_1\bar{N}_2 \\ \hline \bar{M}_2\bar{N}_1 & \bar{C} & \bar{M}_2\bar{N}_2 \end{array} \right)_{i,j} \right\|_2 \tag{58}$$

The $(l+n+1)\times(l+n+1)$ H_2-norm evaluations here require only $l+n+1$ Lyapunov equations to be solved (instead of the $(l+n+p)\times(l+n+m_2)$ equations in the MIMO case represented by (56)), so this expression is preferred.

4.3 Pole Sensitivity Measures

The input-output sensitivity does not explicitly consider the stability of the closed-loop system. To ensure that the implementation is stable, the sensitivity of the poles may be considered. Let $\left(\bar{\lambda}_k\right)_{1\leqslant k\leqslant n_{\mathcal{P}}+n}$ denote the poles of the closed-loop system (the eigenvalues of $\bar{A}$). They are perturbed during the quantization process to $\left(\bar{\lambda}_k^{\dagger}\right)_{1\leqslant k\leqslant n_{\mathcal{P}}+n}$ with

$$\left|\left|\bar{\lambda}_k^{\dagger}\right| - \left|\bar{\lambda}_k\right|\right| \leqslant \sum_{i,j} \Delta_{i,j} \left.\frac{\partial\left|\bar{\lambda}_k^{\dagger}\right|}{\partial \Delta_{i,j}}\right|_{\Delta=0} + o\left(\|\Delta\|_{\max}^2\right). \tag{59}$$

So, we can define the following pole sensitivity measure.

Definition 8 *Consider a controller realization $\mathcal{C} := (Z, l, m_2, n, p_2)$. The closed-loop pole sensitivity measure is defined by*

$$\bar{\Psi} \triangleq \sum_{k=1}^{n_{\mathcal{P}}+n} \left\| \frac{\partial\left|\bar{\lambda}_k\right|}{\partial Z} \times W_Z \right\|_F^2. \tag{60}$$

The following lemma will be required next to evaluate $\bar{\Psi}$.

Lemma 2 *Consider a differentiable function $f : \mathbb{R}^{m\times n} \to \mathbb{C}$, and two matrices $Y \in \mathbb{R}^{m\times n}$ and $X \in \mathbb{R}^{p\times q}$. Let Y_0, Y_1 and Y_2 be constant matrices with appropriate dimensions. Then the following results hold:*

- *if $Y = Y_0 + Y_1 X Y_2$, then*

$$\frac{\partial f(Y)}{\partial X} = {Y_1}^{\top} \frac{\partial f(Y)}{\partial Y} {Y_2}^{\top},$$

- *if* $Y = Y_0 + Y_1 X^{-1} Y_2$, *then*

$$\frac{\partial f(Y)}{\partial X} = -\left(Y_1 X^{-1}\right)^\top \frac{\partial f(Y)}{\partial Y} \left(X^{-1} Y_2\right)^\top.$$

Proof:
See [21]. ■

The measure $\bar{\Psi}$ can be evaluated thanks to the following proposition and lemma.

Proposition 5

$$\frac{\partial \left|\bar{\lambda}_k\right|}{\partial Z} = \bar{M}_1^\top \frac{\partial \left|\bar{\lambda}_k\right|}{\partial \bar{A}} \bar{N}_1^\top \tag{61}$$

where $\bar{M}_1$ *and* $\bar{M}_2$ *are defined in equations* (48) *and* (49).
Proof:
The proof is similar to the one used in Proposition 3, by applying Lemma 2, instead of Lemma 1. ■

Lemma 3 *Let* $M \in \mathbb{R}^{n\times n}$ *be diagonalisable. Let* $(\lambda_k)_{1\leqslant k\leqslant n}$ *be its eigenvalues, and* $(x_k)_{1\leqslant k\leqslant n}$ *the corresponding right eigenvectors. Denote* $M_x \triangleq \left(x_1, x_2, \ldots, x_n\right)$ *and* $M_y = \left(y_1, y_2, \ldots, y_n\right) \triangleq M_x^{-H}$. *Then*

$$\frac{\partial \lambda_k}{\partial M} = y_k^* x_k^\top \quad \forall k = 1, \ldots, n \tag{62}$$

and

$$\frac{\partial \left|\lambda_k\right|}{\partial M} = \frac{1}{\left|\lambda_k\right|} Re\left(\lambda_k^* \frac{\partial \lambda_k}{\partial M}\right) \tag{63}$$

where $\cdot^*$ *denotes the conjugate operation,* $Re(\cdot)$ *the real part and* $\cdot^H$ *the transpose conjugate operator.*
Proof:
See [33]. ■

Remark 7 Similarly to the input-output sensitivity matrix, (41), a *pole sensitivity matrix* can be constructed to evaluate the overall impact of each coefficient. Let $\frac{\delta\left|\bar{\lambda}\right|}{\delta Z}$ denote the pole sensitivity matrix defined by

$$\left(\frac{\delta \left|\bar{\lambda}\right|}{\delta Z}\right)_{i,j} \triangleq \sqrt{\sum_{k=1}^{n_\mathcal{P}+n} \left(\frac{\partial \left|\bar{\lambda}_k\right|}{\partial Z_{i,j}}\right)^2}. \tag{64}$$

It can be computed from

$$\frac{\partial \left|\bar{\lambda}_k\right|}{\partial Z_{i,j}} = \left(\frac{\partial \left|\bar{\lambda}_k\right|}{\partial Z}\right)_{i,j} \tag{65}$$

During the quantization process, Z is perturbed to $Z^\dagger$ and the closed-loop eigenvalues $\left(\bar{\lambda}_k\right)_{1\leqslant k\leqslant n_{\mathcal{P}}+n}$ may be outside the open unit disc. Therefore, it is crucial to know when the FWL error will cause closed-loop instability. Based on this consideration, a stability related measure [4] is defined as:

$$\mu_0(Z) \triangleq \inf_{\Delta} \left\{ \|\Delta\|_{\max} \,/\, \text{realization } Z^\dagger \text{ makes the closed-loop system unstable} \right\} \tag{66}$$

This measure is not directly tractable [4, 32], but can be approached with the following measure.

Definition 9 *Consider a realization $\mathcal{C} := (Z, l, m_2, n, p_2)$. The Pole Sensitivity Stability related Measure (PSSM) of $\mathcal{C}$ is defined by*

$$\mu_1(Z) \triangleq \min_{1\leqslant k\leqslant n_{\mathcal{P}}+n} \frac{1-\left|\bar{\lambda}_k\right|}{\|W_Z\|_F \left\| \frac{\partial \left|\bar{\lambda}_k\right|}{\partial Z} \times W_Z \right\|_F}. \tag{67}$$

This measure evaluates how a perturbation, Δ, of the parameters, Z, can cause instability. It is determined by how close the eigenvalues of $\bar{A}$ are to the unit circle and by how sensitive they are to the controller parameter perturbation.

This measure is an extension to the SIF framework of the sensitivity stability related measure originally defined in the classical state-space framework [21] and can be directly linked to an estimation of the smallest wordlength required for the controller realization to be implemented while preserving the closed-loop stability [34].

4.4 Closed-loop roundoff noise analysis

Complementary to the other two measures, a measure of the roundoff noise is presented next, in the generalized context of the SIF. It extends the measure proposed in [13] to the closed-loop case.

4.4.1 Preliminaries

The first (μ) and second (σ, ψ) order centered-moments of a noise vector $\xi(k)$ are denoted and defined by

$$\mu_\xi \triangleq E\left\{\xi(k)\right\}, \tag{68}$$

$$\psi_\xi \triangleq E\left\{\left(\xi(k)-\mu_\xi\right)\left(\xi(k)-\mu_\xi\right)^\top\right\}, \tag{69}$$

$$\sigma_\xi^2 \triangleq E\left\{\left(\xi(k)-\mu_\xi\right)^\top\left(\xi(k)-\mu_\xi\right)\right\} = tr\left(\psi_\xi\right), \tag{70}$$

where $E\{\cdot\}$ and $tr(\cdot)$ are respectively the *mean* and the *trace* operator.

The following lemma recalls the basic properties of noise transmission through a linear system:

Lemma 4 *Assume the input noise, $U(k)$, to be such that*

$$E\left\{(U(k)-\mu_U)\,(U(k-l)-\mu_U)^\top\right\} = \delta_{0,l}\psi_U \tag{71}$$

where $\delta_{i,j}$ represents the Kronecker delta. Denote by Y the resulting output of the transfer matrix G. If (A,B,C,D) is a state-space realization of G, the first and second order moments of Y are given by:

$$\mu_Y = G(1)\mu_U \tag{72}$$
$$\sigma_Y^2 = tr\left(\psi_U(D^\top D + B^\top W_o B)\right) \tag{73}$$

where $G(1)$ is the steady state gain of G, given by $G(1) = C(I-A)^{-1}B + D$ and W_o is the observability Gramian of G. W_o is the unique solution of the discrete Lyapunov equation

$$W_o = A^\top W_o A + C^\top C \tag{74}$$

Proof:
It is well known that $\sigma_Y^2 = \|G\varphi_U\|_2^2$, with φ_U the square root of ψ_U [26]. The classical formulae linking the H_2 norm to the Gramians is then applied. ■

4.4.2 Roundoff Noise Analysis

Consider the realization $\mathcal{R} := (Z, l, m_2, n, p_2)$. By taking into account the quantization noise after each multiplication, the algorithm given by (3) becomes

$$\begin{aligned}
&\text{[i]} \quad J.T^*(k+1) \leftarrow M.X^*(k) + N.U(k) + \xi_T(k)\\
&\text{[ii]} \quad X^*(k+1) \leftarrow K.T^*(k+1) + P.X^*(k) + Q.U(k) + \xi_X(k)\\
&\text{[iii]} \quad Y^*(k) \leftarrow L.T^*(k+1) + R.X^*(k) + S.U(k) + \xi_Y(k)
\end{aligned} \tag{75}$$

where ξ_T, ξ_X and ξ_Y are respectively the noise sources corrupting T, X and Y (ξ_T is added on $JT(k+1)$, so $J^{-1}\xi_T$ is added on $T(k+1)$).

Noise sources ξ_T, ξ_X and ξ_Y depend on:

- the way the computations are performed, the order of the arithmetic operations, etc.
- the fixed-point representation of the inputs,
- the fixed-point representation of the outputs,
- the fixed-point representation chosen for the states and the intermediate variables,
- the fixed-point representation chosen for the coefficients.

They are modelled as independent white noise, characterized by their first and second order moments.

Remark 8 The quantization or roundoff process can be considered as the addition of a noise, ξ. If ε represents the quantization step, then [31] $\mu_\xi = 0$ and $\sigma_\xi = \varepsilon^2/12$ for roundoff, and $\mu_\xi = \varepsilon/2$ and $\sigma_\xi = \varepsilon^2/12$ for truncation.

The noise is added through the controller and the plant to the output $Z(k)$ of the closed-loop system $\bar{\mathcal{S}}$. Denote the noise added to $Z(k)$ by $\xi'(k)$:

$$\xi'(k) \triangleq Z^*(k) - Z(k) \tag{76}$$

Definition 10 *The Output Noise Power $\bar{P}$ is defined as the power of $\xi'(k)$*

$$\bar{P} \triangleq E\left\{\xi'^{\top}(k)\xi'(k)\right\} \tag{77}$$

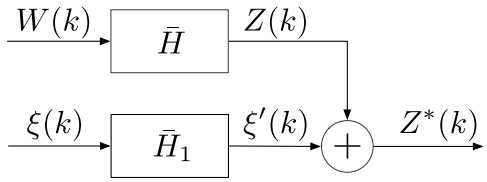

Figure 5: Equivalent system, with noise extracted

Denote by ξ the vector stacking all the noise sources

$$\xi(k) \triangleq \begin{pmatrix} \xi_T(k) \\ \xi_X(k) \\ \xi_Y(k) \end{pmatrix} \tag{78}$$

Proposition 6 *The noise $\xi'(k)$ corresponds to the noise $\xi(k)$ filtered through the transfer function $\bar{H}_1$ defined in eq. (46) (the closed-loop system is then equivalent to the system described in Figure 5). Hence, we get*

$$\bar{P} = tr\left(\psi_\xi\left(\bar{M}_2^{\top}\bar{M}_2 + \bar{M}_1^{\top}\bar{W}_o\bar{M}_1\right)\right) + \mu_{\xi'}^{\top}\mu_{\xi'} \tag{79}$$

where $\mu_{\xi'} = \left(C_Z(I - A_Z)^{-1}\bar{M}_1 + \bar{M}_2\right)\mu_\xi$.
Proof:
If $X_{\mathcal{P}}$ denotes the state of the plant, equation (75) combined with the state-space realization of the plant leads to

$$\begin{cases} \begin{pmatrix} X_{\mathcal{P}} \\ X \end{pmatrix}(k+1) = \bar{A}\begin{pmatrix} X_{\mathcal{P}} \\ X \end{pmatrix}(k) + \bar{B}W(k) + \bar{M}_1\xi(k) \\ Z(k) = \bar{C}\begin{pmatrix} X_{\mathcal{P}} \\ X \end{pmatrix}(k) + \bar{D}W(k) + \bar{M}_2\xi(k) \end{cases} \tag{80}$$

So, $\bar{H}_1$ (cf. eq. (46)) appears explicitly as the transfer linking $\xi(k)$ to $Z(k)$ as stated in the proposition. Therefore, $P = E\left\{\xi'^{\top}(k)\xi'(k)\right\} = \sigma^2_{\xi'} + \mu^{\top}_{\xi'}\mu_{\xi'}$ and Lemma 4 gives the first and second order moment. ■

Remark 9 Equation (79) is a good illustration of the relationship between the work done in the *hardware/software* (HW/SW) community and that done in the *control* community. The former is based on the accurate evaluation of the noise for particular HW/SW fixed-point implementations on various targets (DSP, FPGA) whereas the latter is based on the search for *good* realizations with particular well-conditioned structures. In the first case, only the classical direct form is studied, whereas the actual HW/SW impact is neglected in the second case.

The moments ψ_ξ and μ_ξ depend only on the HW/SW implementation, whereas the other terms ($\bar{A}$, $\bar{C}$, $\bar{M}_1$, $\bar{M}_2$ and $\bar{W}_o$) depend only on the algorithm used.

4.4.3 Roundoff Noise Gain

The *closed-loop roundoff noise gain* is the output noise power in a specific (and simplified) computational scheme: the noise is assumed to appear only after each multiplication (roundoff after multiplication scheme). It is modelled as a zero-mean centered, statistically independent, white noise. Each noise source has the same power σ_0^2 (determined by the wordlength chosen for all the variables and coefficients).

Definition 11 *The closed-loop Roundoff Noise Gain (RNG) is defined as*

$$\bar{G} \triangleq \frac{\bar{P}}{\sigma_0^2} \tag{81}$$

This measure has been studied for the open-loop case by [24, 14, 5] and has been established for classical state-space realizations and some other particular realizations. The particular computational scheme considered gives the moments of ξ_T, ξ_X and ξ_Y: here they depend only on the number of non-trivial parameters in the realization.

Let introduce the matrices d_J to d_S. They are diagonal matrices defined by

$$(d_X)_{i,i} \triangleq \left\{\text{number of non-trivial parameters in the i}^{\text{th}}\text{ row of } X\right\} \tag{82}$$

The trivial parameters considered are 0, 1 and -1 because they do not imply a multiplication.

Step [i] of algorithm (3) is realized as follows (for $i \in \{1, 2, \ldots, l\}$):

$$T_i(k+1) \leftarrow \sum_{j=1}^{n} M_{ij}X_j(k) + \sum_{j=1}^{m} N_{ij}U_j(k) - \sum_{j<i} J_{ij}T_j(k+1) \tag{83}$$

Each multiplication by a non-trivial parameter implies a quantization noise. Since they are independent centered white noise, ψ_{ξ_T} is given by:

$$\psi_{\xi_T} = (d_M + d_N + d_J)\,\sigma_0^2 \tag{84}$$

(J is a lower diagonal matrix with 1 on the diagonal. So the number of non-trivial parameters on the i^{th} row is equal to the number of non-trivial parameters of the i^{th} row restricted to its sub-diagonal part).

In the same way (steps [ii] and [iii]),

$$\psi_{\xi_Y} = (d_L + d_R + d_S)\,\sigma_0^2 \tag{85}$$

$$\psi_{\xi_X} = (d_K + d_P + d_Q)\,\sigma_0^2 \tag{86}$$

Proposition 7 *The RNG is given by*

$$\bar{G} = tr\left(d_Z(\bar{M}_2^\top \bar{M}_2 + \bar{M}_1^\top \bar{W}_o \bar{M}_1)\right) \tag{87}$$

where

$$d_Z = \begin{pmatrix} d_J + d_M + d_N & & \\ & d_K + d_P + d_Q & \\ & & d_L + d_R + d_S \end{pmatrix} \tag{88}$$

(d_Z is also defined by equation (82) *applied on Z)*
Proof:
The noise sources ξ_T, ξ_X and ξ_Y are zero mean centered independent noises so μ_ξ is null and

$$\psi_\xi = \begin{pmatrix} \psi_{\xi_T} & & \\ & \psi_{\xi_X} & \\ & & \psi_{\xi_Y} \end{pmatrix} \tag{89}$$

■

4.5 Comparison to the open-loop measures

In [12, 13], three open-loop measures have been defined. It is worth noting that they are linked to the closed-loop ones:

- the open-loop input-output sensitivity:

$$M_{L_2}^W \triangleq \left\| \frac{\delta H}{\delta Z} \times W_Z \right\|_F^2 \tag{90}$$

  where H is the controller's transfer function (see eq. (7))

- the open-loop pole sensitivity:

$$\Psi \triangleq \left\| \frac{\delta\,|\lambda|}{\delta Z} \times W_Z \right\|_F^2 \tag{91}$$

  where $(\lambda_k)_{1 \leqslant k \leqslant n}$ are the controller's poles.

- and the roundoff noise analysis P defines as the output noise power.

They can be expressed with:

$$\frac{\partial H}{\partial Z} = \left(C_Z(zI_n - A_Z)^{-1}M_1 + M_2\right) \circledast \left(N_2 + N_1(zI_n - A_Z)^{-1}B_Z\right), \tag{92}$$

$$\frac{\partial |\lambda_k|}{\partial Z} = M_1^\top \frac{\partial |\lambda_k|}{\partial A_Z} N_1^\top, \tag{93}$$

$$P = tr\left(\psi_\xi(M_2^\top M_2 + M_1^\top W_o M_1)\right) \tag{94}$$

where

$$M_1 \triangleq \begin{pmatrix} KJ^{-1} & I_n & 0 \end{pmatrix}, \qquad M_2 \triangleq \begin{pmatrix} LJ^{-1} & 0 & I_{p_2} \end{pmatrix}, \tag{95}$$

$$N_1 \triangleq \begin{pmatrix} J^{-1}M \\ I_n \\ 0 \end{pmatrix}, \qquad N_2 \triangleq \begin{pmatrix} J^{-1}N \\ 0 \\ I_{m_2} \end{pmatrix}. \tag{96}$$

The similarities with equations (44), (61) and (77) are obvious.

5 Optimal Design

For the implementation of a digital controller, it is important to choose a realization having low FWL effects. Hence it is of interest to find an optimal realization in a sense to be defined.

Problem 1 *The* global optimal realization problem *is to find the best realization* $\mathcal{R}_{opt}$ *associated with the transfer function* H *according to the criteria* $\mathcal{J}$*:*

$$\mathcal{R}_{opt} = \arg\min_{\mathcal{R}\in\mathfrak{R}_H} \mathcal{J}(\mathcal{R}). \tag{97}$$

Due to the size of $\mathfrak{R}_H$, this problem generally cannot be solved practically. Hence the following problem is introduced to restrict the search to some particular structurations.

Problem 2 () *Consider some structurations* $(\mathcal{S}_i)_{1\leqslant i\leqslant N}$*. The* optimal structured realization problem *is to find the optimal realization* $\mathcal{R}^{\mathcal{S}}_{opt}$*:*

$$\mathcal{R}^{\mathcal{S}}_{opt} = \arg\min_{\substack{\mathcal{R}\in\mathfrak{R}_H^{\mathcal{S}_i} \\ 1\leqslant i\leqslant N}} \mathcal{J}(\mathcal{R}). \tag{98}$$

Since the measure $\mathcal{J}$ could be non-smooth and/or non-convex, the Adaptive Simulated Annealing (ASA) [17, 3] method has been chosen to solve Problem 2. This method has worked well for other optimal realization problems [33].

If the equivalent structured realizations are linked through the similarity transformation of Proposition 2, the computation of the previously defined FWL measures can be improved thanks to the following proposition:

Proposition 8 *If we consider two realizations Z_0 and Z_1 such that:*

$$Z_1 = \mathcal{T}_1 Z_0 \mathcal{T}_2 \tag{99}$$

where

$$\mathcal{T}_1 = \begin{pmatrix} \mathcal{Y} & & \\ & \mathcal{U}^{-1} & \\ & & I_p \end{pmatrix}, \quad \mathcal{T}_2 = \begin{pmatrix} \mathcal{W} & & \\ & \mathcal{U} & \\ & & I_m \end{pmatrix}. \tag{100}$$

then the closed-loop measures of realization Z_1 can be computed from those of Z_0 according to

$$\left(\frac{\delta \bar{H}}{\delta Z}\right)_{i,j}\Bigg|_{Z_1} = \left\| \bar{H}_1\big|_{Z_0} \mathcal{T}_1^{-1} E_{i,j} \mathcal{T}_2^{-1} \bar{H}_2\big|_{Z_0} \right\|_2, \tag{101}$$

$$\frac{\partial \left|\bar{\lambda}_k\right|}{\partial Z}\Bigg|_{Z_1} = \mathcal{T}_1^{-\top} \frac{\partial \left|\bar{\lambda}_k\right|}{\partial Z}\Bigg|_{Z_0} \mathcal{T}_2^{-\top} \tag{102}$$

Proof:
The proof comes directly from

$$\bar{H}_1\big|_{Z_1} = \bar{H}_1\big|_{Z_0} \mathcal{T}_1^{-1}, \qquad \bar{H}_2\big|_{Z_1} = \mathcal{T}_2^{-1} \bar{H}_2\big|_{Z_0}. \tag{103}$$

■

A Matlab toolbox (*FWR Toolbox*[2]) has been specially developed to use the SIF and solve optimal structured realization problems with the previously defined measures.

6 Examples

6.1 Example 1

The first example is taken from [5], pp 236–237. The discrete time system to be controlled is given by

$$A_p = \begin{pmatrix} 3.7156 & -5.4143 & 3.6525 & -0.9642 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \tag{104}$$

$$B_p = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}^\top, \tag{105}$$

$$C_p = \begin{pmatrix} 0.1116 & 0.0043 & 0.1088 & 0.0014 \end{pmatrix} \times 10^{-5}. \tag{106}$$

[2] Available from `http://fwrtoolbox.gforge.inria.fr/`.

Remark 10 All the computations are performed with Matlab double floating-point precision, but the results are quoted only to 4 significants digits (which may be insufficient to characterize the considered system). For each different realization, bold font is used to exhibit non trivial parameters (the weighting matrice W_Z is built accordingly).

It corresponds to the following standard form (see (30))

$$\mathcal{P} := \left(\begin{array}{c|cc} A & B_p & B_p \\ \hline C_p & 0 & 0 \\ C_p & 0 & 0 \end{array}\right) \tag{107}$$

The initial realization of the feedback controller is designed to place the closed-loop poles at

$$\lambda_{1,2} = 0.9844 \pm 0.0357j, \qquad \lambda_{3,4} = 0.9643 \pm 0.0145j, \tag{108}$$
$$\lambda_{5,6} = 0.7152 \pm 0.6348j, \qquad \lambda_{7,8} = 0.3522 \pm 0.2857j. \tag{109}$$

The controller has the following transfer function

$$H : z \mapsto \frac{38252z^3 - 101878z^2 + 91135z - 27230}{z^4 - 2.3166z^3 + 2.1662z^2 - 0.96455z + 0.17565} \tag{110}$$

Let us consider different realizations for this controller. The realizations, Z_1 to Z_{11}, are described below. The values of the measures are shown in Table 1. The realizations and corresponding sensitivity matrices, $\frac{\delta \bar{H}}{\delta Z}$ and $\frac{\delta |\bar{\lambda}|}{\delta Z}$, are given in the appendix. Note that only the bold values shown in the realizations are considered, via the weighting matrix W_Z.

State-space realizations

Z_1: Canonical form (corresponds to Direct Form II). This realization has the following results

$$\bar{M}^W_{L_2} = 1.9046e{+}7, \quad \bar{\Psi} = 3.3562e{+}7, \quad \mu_1 = 1.8065e{-}6, \quad \bar{G} = 1.186e{+}6 \tag{111}$$

Z_2: The *internally balanced* state-space realization is often considered as a low sensitivity realization ([5] shows that the balanced realizations minimizes the L_1/L_2 sensitivity measure). It has the following measure values

$$\bar{M}^W_{L_2} = 3.6427e{+}5, \quad \bar{\Psi} = 6.5007e{+}5, \quad \mu_1 = 7.4933e{-}6, \quad \bar{G} = 365.82. \tag{112}$$

Despite it being fully parametrized (24 parameters), its overall sensitivity is lower than the canonical form.

Z_3: With the similarity

$$\mathcal{T}_1 = \begin{pmatrix} \cdot & & \\ & \mathcal{U}^{-1} & \\ & & 1 \end{pmatrix}, \quad \mathcal{T}_2 = \begin{pmatrix} \cdot & & \\ & \mathcal{U} & \\ & & 1 \end{pmatrix} \tag{113}$$

it is possible to consider all state-space equivalent realizations, and find the $\bar{M}^W_{L_2}$-optimal state-space realization Z_3. Its closed-loop transfer function sensitivity measure is $\bar{M}^W_{L_2} = 1526.7$ and is much lower than other state space realizations.

Z_4: It is also possible to consider the $\bar{\Psi}$-optimal state-space realization. Then $\bar{\Psi} = 2742.5$.

Z_5: $\bar{G}$-optimal state-space Z_5. Here, $\bar{G}$ is very low: $\bar{G} = 0.0032261$, but the other measure are quite poor:

$$\bar{M}^W_{L_2} = 1.9474e{+}13, \qquad \bar{\Psi} = 1.2294e{+}13, \qquad \mu_1 = 1.7244e{-}9. \tag{114}$$

Even if the goal of this paper is not multi-objective optimal realization, it is interesting to look for a realization that is *good enough* for the three measures $\bar{M}^W_{L_2}$, $\bar{\Psi}$ and $\bar{G}$. Let us denote

$$\bar{TO}(Z) \triangleq \frac{\bar{M}^W_{L_2}(Z)}{\bar{M}^{W\ opt}_{L_2}} + \frac{\bar{\Psi}(Z)}{\bar{\Psi}^{opt}} + \frac{\bar{G}(Z)}{\bar{G}^{opt}} \tag{115}$$

where $\bar{M}^{W\ opt}_{L_2}$ is the optimal transfer function sensitivity value ($\bar{M}^{W\ opt}_{L_2} = \bar{M}^W_{L_2}(Z_3)$), $\bar{\Psi}^{opt}$ the optimal value for the pole sensitivity ($\bar{\Psi}^{opt} = \bar{\Psi}(Z_4)$) and $\bar{G}^{opt}$ the optimal roundoff noise gain value ($\bar{G}^{opt} = \bar{G}(Z_5)$).

Remark 11 This *tradeoff* measure is defined for this example and this structuration (state-space). Clearly, it is lower bounded by 3.

Z_6: *tradeoff*-optimal state-space Z_6. With this measure, we aim to have a realization that simultaneously has low transfer function sensitivity, pole sensitivity and roundoff noise gain. The *tradeoff* measure is quite low ($\bar{TO} = 6.0078$), and the corresponding measures are:

$$\bar{M}^W_{L_2} = 2869.6, \qquad \bar{\Psi} = 4537.1, \qquad \mu_1 = 9.2351e{-}5, \qquad \bar{G} = 0.0079809. \tag{116}$$

ρ **Direct Forms II transposed** The realization (25) is considered with various values for $(\gamma_i)_{1\leqslant i\leqslant n}$. Δ is chosen to be 2^{-3}. Since there is no possibility here to use similarity on Z like that proposed in Proposition 2, the realization matrix Z cannot be built from another Z matrix : for $(\gamma_i)_{1\leqslant i\leqslant n}$ given, the parameters $(\alpha_i)_{1\leqslant i\leqslant n}$ and $(\beta_i)_{0\leqslant i\leqslant n}$ have to be rebuilt from (20).

Z_7: with $\gamma = \begin{pmatrix}1 & 1 & 1 & 1\end{pmatrix}^\top$, the Direct Form II with the δ-operator is obtained.

Z_8: $M_{L_2^W}$-optimal ρDFIIt. The optimization gives

$$\gamma = \begin{pmatrix}0.29758 & 0.99939 & 0.99953 & 0.99977\end{pmatrix}^\top \tag{117}$$

Z_9: $\bar{\Psi}$-optimal ρDFIIt. The optimization gives

$$\gamma = \begin{pmatrix} 0.35114 & 0.30858 & 0.66309 & 0.99856 \end{pmatrix}^\top \tag{118}$$

Z_{10}: $\bar{G}$-optimal ρDFIIt. The optimization gives

$$\gamma = \begin{pmatrix} 0.93207 & 0.99335 & 0.99863 & 0.99963 \end{pmatrix}^\top \tag{119}$$

Z_{11}: It is here also possible to apply a new *tradeoff* measure, like the one in equation (115) (with new $\bar{M}_{L_2}^{W\ opt}$, $\bar{\Psi}^{opt}$ and $\bar{G}^{opt}$ values). The $\bar{TO}$-optimal realization (eq. (138)) is obtained with

$$\gamma = \begin{pmatrix} 0.99744 & 0.41349 & 0.8646 & 0.99346 \end{pmatrix}^\top \tag{120}$$

and $\bar{TO} = 3.5597$.

Table 1 gives all the measure values for the realization Z_1 to Z_{11}. Realizations Z_6 and Z_{11} are interesting, low sensitivity, low roundoff noise, realizations. Moreover Z_{11} requires fewer operations (11 additions and 16 multiplications) than Z_6. These results are case dependent and some controllers may be less sensitive in state-space forms than in ρDFIIt form.

Table 1: Example 1: FWL measures for different realizations

	$\bar{M}_{L_2}^{W}$	$\bar{\Psi}$	μ_1	$\bar{G}$	$\bar{TO}$	**Nb. op.**
Z_1	1.9046e+7	3.3562e+7	1.8065e−6	1.186e+6	3.6764e+8	$7 + 8\times$
Z_2	3.6427e+5	6.5007e+5	7.4933e−6	3.6582e+2	1.1387e+5	$19 + 24\times$
Z_3	1.5267e+3	1.6689e+4	1.167e−4	1.7455e+2	5.4111e+4	$19 + 24\times$
Z_4	1.6272e+3	2.7425e+3	1.189e−4	1.1778e+2	3.6512e+4	$19 + 24\times$
Z_5	1.9474e+13	1.2294e+13	1.7244e−9	3.2261e−3	1.7239e+10	$19 + 24\times$
Z_6	2.8696e+3	4.5371e+3	9.2351e−5	7.9809e−3	6.0078e+0	$19 + 24\times$
Z_7	1.5342e−2	8.1051e−2	6.6047e−2	2.8082e−8	4.5466e+0	$11 + 12\times$
Z_8	1.5341e−2	8.089e−2	6.6045e−2	4.217e−8	4.8783e+0	$11 + 16\times$
Z_9	1.1388e−1	2.8203e−2	6.6159e−2	3.7783e−6	9.8937e+1	$11 + 16\times$
Z_{10}	1.5342e−2	8.0015e−2	6.6052e−2	4.1742e−8	4.8371e+0	$11 + 16\times$
Z_{11}	1.6065e−2	3.8802e−2	6.0413e−2	4.7451e−8	3.5597e+0	$11 + 16\times$

The pseudocode algorithms associated with realizations Z_6 and Z_{11} are given by Algorithms 1 and 3 listed in the appendix. It is assumed that these realizations are performed on a fixed-point 16-bit processor (the additions are 32 bits, without guard bits for the additions) and the input is in the interval $[-10, 10]$ (so 11 bits are given for the fractional part). Due

to the gain of the controller, the output has -5 bits for the fractional part (the integer value coding for the output must be multiplied by 2^6 to obtain the real value). The binary point position is adjust for each intermediate variable, state and coefficient. So the fixed-point algorithms of realizations Z_6 and Z_{11} are given by Algorithms 2 and 4.

6.2 Example 2

The second numerical example is the active control of longitudinal vehicle oscillations studied in [20]. One significant aspect of vehicle driveability is the attenuation of the first torsional mode (resonance in the elastic parts) which produces unpleasant (0 to 10 Hz) longitudinal oscillations of the vehicle, known as shuffle. They can be reduced by means of a controller acting on the engine torque.

The discretized model $P(z)$ of the power train is given by (141) and (142), and a discrete-time realization of the controller is given by (142) and (143) – this being an H_∞ balanced realization.

The different forms studied here are :

Z_4: direct form II

Z_5: $\bar{M}_{L_2}^W$-optimal state-space

Z_6: $\bar{\Psi}$-optimal classical state-space

Z_7: Direct form IIt with δ-operator (equivalent to ρDFIIt form with $\gamma_i = 1$ and $\Delta_i = 2^{-5}$)

Z_8: $\bar{M}_{L_2}^W$-optimal ρDFIIt form ($\Delta_i = 2^{-5}$)

Z_9: $\bar{\Psi}$-optimal ρDFIIt form ($\Delta_i = 2^{-5}$)

Table 2 shows the different sensitivity values. The optimal realization Z_8 is obtained with $\gamma_i = 1$ (so $Z_8 = Z_7$), and Z_9 corresponds to

$$\gamma = \begin{pmatrix} 0.6261617 \\ 0.3288406 \\ 0.04442233 \\ 0.5848309 \\ 0.696381 \\ 0.7397787 \\ 0.6405012 \\ 0.9255434 \\ 0.9729877 \\ 0.9848002 \end{pmatrix} \tag{121}$$

Table 2: Example 2: Closed-loop sensitivities and computational cost for different realizations

realization	$\bar{M}^{W}_{L_2}$	$\bar{\Psi}$	**Nb. operations**
Z_4	$2.8863e{+}23$	$1.7693e{+}16$	$20 + \ 21\times$
Z_5	$2.3167e{+}5$	$1.8680e{+}6$	$110 + \ 121\times$
Z_6	$6.4165e{+}4$	$8.1927e{+}6$	$110 + \ 121\times$
Z_7	$8.7491e{-}2$	$2.6161e{+}5$	$30 + \ 31\times$
Z_8	$8.7491e{-}2$	$2.6161e{+}5$	$30 + \ 31\times$
Z_9	$1.5759e{+}5$	$8.0501e{+}3$	$30 + \ 41\times$

7 Conclusions

The Specialized Implicit Form is a powerful tool for filter and controller implementation modelling. It provides a macroscopic description of the algorithm to be implemented, in the context of embedded systems. More general than previous forms, it allows, in a unified framework, the analysis and design of particular realizations of linear controllers. Different measures can give insight on the quality of a given realization: input-output sensitivity, pole sensitivity, roundoff noise gain, amount of computation, etc. All have been defined in the new context of the SIF. Some of them are worked out in a efficient way through the use of Gramians and Lyapunov equations.

The notion of equivalence between realizations has been defined, using the inclusion principle. As a consequence, a large variety of realizations, not necessarily of the same order, may be compared. Some optimizations are computationally tractable, by restricting the class of equivalent realizations to specific subclasses or structures. This has been tested in the case of classical state-space realizations, with δ-structures, observer-based realizations, etc. The sparse realization proposed recently in [23] has also been examined.

There are numerous areas for future work. First, it would be of practical interest to make use of the SIF to propose some practical realizations that are generically good (sparse and faithful) in a given context. Second is the modelling of internal delay, this being both computational delay and communication time delay, for example when the controller algorithm has to be split on different processors. Third is to take more precisely into account the hardware/software target, so linking the present work more deeply with what is done in the hardware/software community. Last but not least, improving the optimization process (cheap evaluation of the measures, choice and tuning of the optimization solver, distance evaluation to the optimal optimum) is still an important challenge, although the developed Matlab toolbox, the *FWR Toolbox*, has been able to provide interesting results in different situations.

References

[1] J. Aplevich. *Implicit Linear Systems*. Springer-Verlag, 1991.

[2] L. Bakule, J. Rodellar, and J. Rossell. Structure of expansion-contraction matrices in the inclusion principle for dynamic systems. *SIAM Matrix Anal. Appl.*, 21(4):1136–1155, 2000.

[3] S. Chen and B.L. Luk. Adaptive simulated annealing for optimization in signal processing applications. *Signal Processing*, 79:117–128, 1999.

[4] I. Fialho and T. Georgiou. On stability and performance of sampled-data systems subject to wordlength constraint'. *IEEE Trans. Automatic Control*, 39(12):2476–2481, December 1994.

[5] M. Gevers and G. Li. *Parametrizations in Control, Estimation and Filtering Probems*. Springer-Verlag, 1993.

[6] R. Goodall. Perspectives on processing for real-time control. *Annual Reviews in Control*, 25:123–131, 2001.

[7] J. Hao and G. Li. An efficient structure for finite precision implementation of digital systems. In *Information, Communications and Signal Processing, 2005 Fifth International Conference on*, pages 564–568, 2005.

[8] T. Hilaire. *Analyse et synthèse de l'implémentation de lois de contrôle-commande en précision finie (Étude dans le cadre des applications automobiles sur calculateur embarquée)*. PhD thesis, Université de Nantes, June 2006.

[9] T. Hilaire, P. Chevrel, and Y. Trinquet. Designing low parametric sensitivity FWL realizations of LTI controllers/filters within the implicit state-space framework. In *Proc. of the 44th IEEE Conference on Decision and Control and the European Control Conference (CDC-ECC'05)*, pages 5192–5197, December 2005.

[10] T. Hilaire, P. Chevrel, and Y. Trinquet. Implicit state-space representation : a unifying framework for FWL implementation of LTI systems. In P. Piztek, editor, *Proc. of the 16th IFAC World Congress*. Elsevier, July 2005.

[11] T. Hilaire, P. Chevrel, and J. Whidborne. Low parametric closed-loop sensitivity realizations using fixed-point and floating-point arithmetic. In *Proc. European Control Conference (ECC'07)*, July 2007.

[12] T. Hilaire, P. Chevrel, and J.F. Whidborne. A unifying framework for finite wordlength realizations. *IEEE Trans. on Circuits and Systems*, 8(54), August 2007.

[13] T. Hilaire, D. Ménard, and O. Sentieys. Roundoff noise analysis of finite wordlength realizations with the implicit state-space framework. In *15th European Signal Processing Conference (EUSIPOC'07)*, September 2007.

[14] S.Y. Hwang. Minimum uncorrelated unit noise in state-space digital filtering. *IEEE Trans. on Acoust., Speech, and Signal Processing*, 25(4):273–281, August 1977.

[15] M. Ikeda and D. Šiljak. Overlapping decompositions, expansions, and contractions of dynamic systems. *Large Scale Syst.*, 1:29–38, 1980.

[16] M. Ikeda, D. Šiljak, and D. White. An inclusion principle for dynamic systems. *IEEE Trans. Automatic Control*, 29(3):244–249, March 1984.

[17] L. Ingber. Adaptive simulated annealing (ASA): Lessons learned. *Control and Cybernetics*, 25(1):33–54, 1996.

[18] R.S.H. Istepanian and J.F. Whidborne, editors. *Digital Controller Implementation and Fragility: A Modern Perspective*. Springer-Verlag, London, UK, September 2001.

[19] H.-J. Ko and W.-S. Yu. Improved eigenvalue sensitivity for finite-precision digital controller realisations via orthogonal Hermitian transform. *IEE Proc. Control Theory and Appl.*, 150(4):365–375, 2003.

[20] D. Lefebvre, P. Chevrel, and S. Richard. An H_∞ based control design methodology dedicated to the active control of longitudinal oscillations. *IEEE Trans. on Control Systems Technology*, 11(6):948–956, November 2003.

[21] G. Li. On the structure of digital controllers with finite word length consideration. *IEEE Trans. on Autom. Control*, 43(5):689–693, May 1998.

[22] G. Li. A polynomial-operator-based dfiit structure for iir filters. *IEEE Trans. on Circuits and Systems-II*, 51(3):147–151, March 2004.

[23] G. Li and Z. Zhao. On the generalized DFIIt structure and its state-space realization in digital filter implementation. *IEEE Trans. on Circuits and Systems*, 51(4):769–778, April 2004.

[24] C. Mullis and R. Roberts. Synthesis of minimum roundoff noise fixed point digital filters. In *IEEE Transactions on Circuits and Systems*, volume CAS-23, September 1976.

[25] M. Palaniswami and G. Feng. Digital estimation and control with a new discrete time operator. In *Proc. 30th IEEE Conf. Decision Contr.*, pages 1631–1632, Brighton, U.K., December 1991.

[26] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. Mc Graw Hill, 1991.

[27] L. Thiele. Design of sensitivity and round-off noise optimal state-space discrete systems. *Int. J. Circuit Theory Appl.*, 12:39–46, 1984.

[28] D. Šiljak. *Decentralized Control of Complex Systems*. Academic Press, 1991.

[29] J.F. Whidborne, R. Istepanian, and J. Wu. Reduction of controller fragility by pole sensitivity minimization. *IEEE Trans. Automatic Control*, 46:320–325, 2001.

[30] J.F. Whidborne, J. Wu, and R.H. Istepanian. Finite word length stability issues in an ℓ_1 framework. *Int. J. Control*, 73(2):166–176, 2000.

[31] B. Widrow. Statistical analysis of amplitude quantized sampled-data systems. *Trans AIEE*, 2(79):555–568, 1960.

[32] J. Wu and S. Chen. *Unsolved Problems in Mathematical Systems and Control Theory*, chapter Stable controller coefficient perturbation in floating point implementation, pages 280–284. Princeton University Press, 2004.

[33] J. Wu, S. Chen, G. Li, R. Istepanian, and J. Chu. An improved closed-loop stability related measure for finite-precision digital controller realizations. *IEEE Trans. Automatic Control*, 46(7):1162–1166, 2001.

[34] J. Wu, S. Chen, J.F. Whidborne, and J. Chu. A unified closed-loop stability measure for finite-precision digital controller realizations implemented in different representation schemes. *IEEE Trans. Automatic Control*, 48(5):816–823, May 2003.

[35] K. Zhou, J. Doyle, and K. Glover. *Robust and Optimal Control*. Prentice-Hall, 1996.

A Algorithms and numerical values

Input: u : real
Output: y : real
Data: xn : array of four reals
Data: xnp : array of four reals
Data: Acc : real
begin
  // *compute* $xnp(1)$
  $Acc \leftarrow xn(1) * 1.0056699573;$
  $Acc \leftarrow Acc + xn(2) * -0.3855253273;$
  $Acc \leftarrow Acc + xn(3) * 0.7882084769;$
  $Acc \leftarrow Acc + xn(4) * -0.8602211557;$
  $xnp(1) \leftarrow Acc + u * -1991.2978135292;$
  // *compute* $xnp(2)$
  $Acc \leftarrow xn(1) * -1.7060282729;$
  $Acc \leftarrow Acc + xn(2) * 1.1129704773;$
  $Acc \leftarrow Acc + xn(3) * 0.6255751647;$
  $Acc \leftarrow Acc + xn(4) * -3.4333411367;$
  $xnp(1) \leftarrow Acc + u * 5980.9414091468;$
  // *compute* $xnp(3)$
  $Acc \leftarrow xn(1) * -0.8063580681;$
  $Acc \leftarrow Acc + xn(2) * 0.3468387941;$
  $Acc \leftarrow Acc + xn(3) * 0.5800952206;$
  $Acc \leftarrow Acc + xn(4) * -0.9426058134;$
  $xnp(3) \leftarrow Acc + u * 4482.5598405197;$
  // *compute* $xnp(4)$
  $Acc \leftarrow xn(1) * -2.5973181092;$
  $Acc \leftarrow Acc + xn(2) * 1.5009691911;$
  $Acc \leftarrow Acc + xn(3) * -1.9422913020;$
  $Acc \leftarrow Acc + xn(4) * -0.3821356552;$
  $xnp(4) \leftarrow Acc + u * 15599.2014809957;$
  // *compute the output*
  $Acc \leftarrow xn(1) * 1.3425518386;$
  $Acc \leftarrow Acc + xn(2) * -0.0635813666;$
  $Acc \leftarrow Acc + xn(3) * -0.5530485340;$
  $y \leftarrow Acc + xn(4) * 2.8068277711;$
  // *save the states*
  $xn \leftarrow xnp$
end

Algorithm 1: Realization Z_6

Input: u : 16 bits integer
Output: y : 16 bits integer
Data: xn : array of four 16 bits integers
Data: xnp : array of four 16 bits integers
Data: Acc : 32 bits integer
begin
  // *compute* $xnp(1)$
  $Acc \leftarrow xn(1) * 16477;$
  $Acc \leftarrow Acc + xn(2) * -12633;$
  $Acc \leftarrow Acc + xn(3) * 6457;$
  $Acc \leftarrow Acc + xn(4) * -7047;$
  $Acc \leftarrow Acc + u * -498;$
  $xnp(1) \leftarrow Acc >> 14;$
  // *compute* $xnp(2)$
  $Acc \leftarrow xn(1) * -13976;$
  $Acc \leftarrow Acc + xn(2) * 18235;$
  $Acc \leftarrow Acc + xn(3) * 2562;$
  $Acc \leftarrow Acc + xn(4) * -14063;$
  $Acc = Acc + u * 748;$
  $xnp(2) \leftarrow Acc >> 14;$
  // *compute* $xnp(3)$
  $Acc \leftarrow xn(1) * -26423;$
  $Acc \leftarrow Acc + xn(2) * 22730;$
  $Acc \leftarrow Acc + xn(3) * 9504;$
  $Acc \leftarrow Acc + xn(4) * -15444;$
  $Acc \leftarrow Acc + u * 2241;$
  $xnp(3) \leftarrow Acc >> 14;$
  // *compute* $xnp(4)$
  $Acc \leftarrow xn(1) * -21277;$
  $Acc \leftarrow Acc + xn(2) * 24592;$
  $Acc \leftarrow Acc + xn(3) * -7956;$
  $Acc \leftarrow Acc + xn(4) * -1565;$
  $Acc \leftarrow Acc + u * 1950;$
  $xnp(4) \leftarrow Acc >> 12;$
  // *compute the output*
  $Acc \leftarrow xn(1) * 21996;$
  $Acc \leftarrow Acc + xn(2) * -2083;$
  $Acc \leftarrow Acc + xn(3) * -4531;$
  $Acc \leftarrow Acc + xn(4) * 22994;$
  $y \leftarrow Acc >> 15;$
  // *save the states*
  $xn \leftarrow xnp$
end

Algorithm 2: Fixed-point algorithm of realization Z_6

Input: u : real
Output: y : real
Data: xn : array of four reals
Data: Acc : real
Data: T : array of four reals
begin
  // *Intermediate variables*
  $T(1) \leftarrow xn(1) * 0.125;$
  $T(2) \leftarrow xn(2) * 0.125;$
  $T(3) \leftarrow xn(3) * 0.125;$
  $T(4) \leftarrow xn(4) * 0.125;$
  // *compute* $xn(1)$
  $Acc \leftarrow T(1) * -8.5940609251;$
  $Acc \leftarrow Acc + T(2);$
  $Acc \leftarrow Acc + xn(1) * 0.9974440349;$
  $xn(1) \leftarrow Acc + u * 306012.0144582504;$
  // *compute* $xn(2)$
  $Acc \leftarrow T(1) * -35.2839059945;$
  $Acc \leftarrow Acc + T(3);$
  $Acc \leftarrow Acc + xn(2) * 0.4134893631;$
  $xn(2) \leftarrow Acc + u * -660870.6659178101;$
  // *compute* $xn(3)$
  $Acc \leftarrow T(1) * -201.7634931054;$
  $Acc \leftarrow Acc + T(4);$
  $Acc \leftarrow Acc + xn(3) * 0.9864594697;$
  $xn(3) \leftarrow Acc + u * 966164.3351972550;$
  // *compute* $xn(4)$
  $Acc \leftarrow T(1) * -237.4643508571;$
  $Acc \leftarrow Acc + xn(4) * 0.9934647479;$
  $xn(4) \leftarrow Acc + u * 1086873.2436256856;$
  // *compute the output*
  $y \leftarrow T(1);$
end

Algorithm 3: Realization Z_{11}

Input: u : 16 bits integer
Output: y : 16 bits integer
Data: xn : array of four 16 bits integers
Data: Acc : 32 bits integer
Data: T : array of four 16 bits integers
begin
  // *Intermediate variables*
  $T \leftarrow xn;$
  // *compute* $xn(1)$
  $Acc \leftarrow T(1) * -17601;$
  $Acc \leftarrow Acc + T(2) << 13;$
  $Acc \leftarrow Acc + xn(1) * 16342;$
  $Acc \leftarrow Acc + u * 4781;$
  $xn(1) \leftarrow Acc >> 14;$
  // *compute* $xn(2)$
  $Acc \leftarrow T(1) * -18065;$
  $Acc \leftarrow Acc + T(3) << 13;$
  $Acc \leftarrow Acc + xn(2) * 6775;$
  $Acc \leftarrow Acc + u * -2582;$
  $xn(2) \leftarrow Acc >> 14;$
  // *compute* $xn(3)$
  $Acc \leftarrow T(1) * -25826;$
  $Acc \leftarrow Acc + T(4) << 12;$
  $Acc \leftarrow Acc + xn(3) * 16162;$
  $Acc \leftarrow Acc + u * 944;$
  $xn(3) \leftarrow Acc >> 14;$
  // *compute* $xn(4)$
  $Acc \leftarrow T(1) * -30395;$
  $Acc \leftarrow Acc + xn(4) * 32554;$
  $Acc \leftarrow Acc + u * 1061;$
  $xn(4) \leftarrow Acc >> 15;$
  // *compute the output*
  $y \leftarrow T(1);$
end

Algorithm 4: Fixed-point algorithm of realization Z_{11}

$$Z_1 = \left(\begin{array}{cccc|c} & & & & \\ \hline 0 & 0 & 0 & \mathbf{-0.17565} & 1 \\ 1 & 0 & 0 & \mathbf{0.96455} & 0 \\ 0 & 1 & 0 & \mathbf{-2.1662} & 0 \\ 0 & 0 & 1 & \mathbf{2.3166} & 0 \\ \hline \mathbf{38252} & \mathbf{-13264} & \mathbf{-22452} & \mathbf{-13615} & 0 \end{array}\right) \tag{122}$$

$$Z_2 = \left(\begin{array}{cccc|c} & & & & \\ \hline \mathbf{0.11188} & \mathbf{-0.54082} & \mathbf{0.19539} & \mathbf{-0.053116} & \mathbf{203.18} \\ \mathbf{0.54082} & \mathbf{0.72159} & \mathbf{0.1647} & \mathbf{-0.034978} & \mathbf{63.57} \\ \mathbf{0.19539} & \mathbf{-0.1647} & \mathbf{0.76428} & \mathbf{0.12977} & \mathbf{-32.042} \\ \mathbf{0.053116} & \mathbf{-0.034978} & \mathbf{-0.12977} & \mathbf{0.71885} & \mathbf{-4.1143} \\ \hline \mathbf{203.18} & \mathbf{-63.57} & \mathbf{-32.042} & \mathbf{4.1143} & 0 \end{array}\right) \tag{123}$$

$$\left.\frac{\delta \bar{H}}{\delta Z}\right|_{Z_1} = \left(\begin{array}{cccc|c} & & & & \\ \hline 57.957 & 424.43 & 658.23 & \mathbf{499.8} & 30.319 \\ 429.23 & 3142.7 & 4873.9 & \mathbf{3700.8} & 224.5 \\ 260.34 & 1906.1 & 2956.2 & \mathbf{2244.6} & 136.16 \\ 28.813 & 210.65 & 326.73 & \mathbf{248.07} & 15.049 \\ \hline \mathbf{0.012735} & \mathbf{0.093245} & \mathbf{0.14461} & \mathbf{0.1098} & 0.0066609 \end{array}\right) \tag{124}$$

$$\left.\frac{\delta\,|\bar{\lambda}|}{\delta Z}\right|_{Z_1} = \left(\begin{array}{cccc|c} & & & & \\ \hline 72.508 & 543.22 & 841.66 & \mathbf{642.27} & 39.27 \\ 554.92 & 4148.8 & 6429.3 & \mathbf{4904} & 298.92 \\ 344.53 & 2546.9 & 3950.5 & \mathbf{3006.7} & 180.19 \\ 20.963 & 200.69 & 305.4 & \mathbf{242.93} & 19.482 \\ \hline \mathbf{0.01643} & \mathbf{0.12276} & \mathbf{0.19025} & \mathbf{0.1451} & 0.0088362 \end{array}\right) \tag{125}$$

$$\left.\frac{\delta \bar{H}}{\delta Z}\right|_{Z_2} = \left(\begin{array}{cccc|c} & & & & \\ \hline \mathbf{19.822} & \mathbf{75.488} & \mathbf{73.165} & \mathbf{22.776} & \mathbf{0.36336} \\ \mathbf{75.488} & \mathbf{287.48} & \mathbf{278.64} & \mathbf{86.738} & \mathbf{1.3838} \\ \mathbf{73.165} & \mathbf{278.64} & \mathbf{270.06} & \mathbf{84.069} & \mathbf{1.3412} \\ \mathbf{22.776} & \mathbf{86.738} & \mathbf{84.069} & \mathbf{26.17} & \mathbf{0.41751} \\ \hline \mathbf{0.36336} & \mathbf{1.3838} & \mathbf{1.3412} & \mathbf{0.41751} & 0.0066609 \end{array}\right) \tag{126}$$

$$\left.\frac{\delta\,|\bar{\lambda}|}{\delta Z}\right|_{Z_2} = \left(\begin{array}{cccc|c} & & & & \\ \hline \mathbf{25.368} & \mathbf{99.068} & \mathbf{97.819} & \mathbf{30.298} & \mathbf{0.47514} \\ \mathbf{99.068} & \mathbf{384.42} & \mathbf{375.49} & \mathbf{114.09} & \mathbf{1.8431} \\ \mathbf{97.819} & \mathbf{375.49} & \mathbf{360.23} & \mathbf{105.84} & \mathbf{1.7991} \\ \mathbf{30.298} & \mathbf{114.09} & \mathbf{105.84} & \mathbf{29.048} & \mathbf{0.54599} \\ \hline \mathbf{0.47514} & \mathbf{1.8431} & \mathbf{1.7991} & \mathbf{0.54599} & 0.0088362 \end{array}\right) \tag{127}$$

$$Z_3 = \left(\begin{array}{cccc|c} & & & & \\ \hline \mathbf{3.0771} & \mathbf{1.9943} & \mathbf{-3.5223} & \mathbf{-0.81099} & \mathbf{-8.6995} \\ \mathbf{19.018} & \mathbf{17.794} & \mathbf{-28.317} & \mathbf{-4.7792} & \mathbf{-14.709} \\ \mathbf{15.651} & \mathbf{13.987} & \mathbf{-22.86} & \mathbf{-4.4711} & \mathbf{-24.353} \\ \mathbf{-11.38} & \mathbf{-10.264} & \mathbf{17.463} & \mathbf{4.3055} & \mathbf{19.502} \\ \hline \mathbf{3953.9} & \mathbf{3517.5} & \mathbf{-5956.1} & \mathbf{-1059.4} & 0 \end{array}\right) \tag{128}$$

$$\left.\frac{\delta \bar{H}}{\delta Z}\right|_{Z_3} = \left(\begin{array}{cccc|c} & & & & \\ \hline \mathbf{5.1146} & \mathbf{7.8587} & \mathbf{4.5637} & \mathbf{9.9049} & \mathbf{0.46308} \\ \mathbf{7.7105} & \mathbf{10.124} & \mathbf{5.6179} & \mathbf{14.825} & \mathbf{0.7032} \\ \mathbf{4.2952} & \mathbf{5.9477} & \mathbf{3.3588} & \mathbf{8.2768} & \mathbf{0.39087} \\ \mathbf{10.161} & \mathbf{14.616} & \mathbf{8.3462} & \mathbf{19.613} & \mathbf{0.92301} \\ \hline \mathbf{0.072972} & \mathbf{0.093663} & \mathbf{0.051546} & \mathbf{0.14019} & 0.0066609 \end{array}\right) \tag{129}$$

$$Z_4 = \left(\begin{array}{ccccc:c} & & & & & \\ \hdashline 2.1976 & 2.225 & 1.4698 & -0.6568 & -77.48 \\ 0.18131 & -0.82788 & -1.5695 & -0.4138 & 69.498 \\ -0.95285 & 1.0322 & 2.2218 & 0.88142 & -45.666 \\ 2.5862 & -0.54545 & -1.6235 & -1.2749 & 42.167 \\ \hdashline -394.63 & 40.523 & 200.59 & 332.48 & 0 \end{array}\right) \tag{130}$$

$$\left.\frac{\delta\,|\bar{\lambda}|}{\delta Z}\right|_{Z_4} = \left(\begin{array}{ccccc:c} & & & & & \\ \hdashline 3.9182 & 5.266 & 13.221 & 7.6474 & 0.40421 \\ 10.728 & 6.3818 & 9.0377 & 19.835 & 0.041826 \\ 11.711 & 19.409 & 11.55 & 18.272 & 0.8029 \\ 2.949 & 13.603 & 24.317 & 8.2111 & 0.87066 \\ \hdashline 0.0066022 & 0.14819 & 0.22917 & 0.045837 & 0.0088362 \end{array}\right) \tag{131}$$

$$Z_5 = \left(\begin{array}{ccccc:c} & & & & & \\ \hdashline 26860 & 1.1171e{+}5 & -64054 & 16716 & 6.2454e{+}8 \\ 3731.3 & 15520 & -8898.5 & 2322.2 & 8.4763e{+}7 \\ 23883 & 99334 & -56955 & 14864 & 5.5625e{+}8 \\ 23421 & 97413 & -55854 & 14577 & 5.612e{+}8 \\ \hdashline -0.18915 & -0.78675 & 0.4511 & -0.11772 & 0 \end{array}\right) \tag{132}$$

$$Z_6 = \left(\begin{array}{ccccc:c} & & & & & \\ \hdashline 1.0057 & -0.38553 & 0.78821 & -0.86022 & -1991.3 \\ -1.706 & 1.113 & 0.62558 & -3.4333 & 5980.9 \\ -0.80636 & 0.34684 & 0.5801 & -0.94261 & 4482.6 \\ -2.5973 & 1.501 & -1.9423 & -0.38214 & 15599 \\ \hdashline 1.3426 & -0.063581 & -0.55305 & 2.8068 & 0 \end{array}\right) \tag{133}$$

$$Z_7 = \left(\begin{array}{cccc:cccc:c} -1 & 0 & 0 & 0 & 0.125 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0.125 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0.125 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0.125 & 0 \\ \hdashline -13.467 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 3.0601e{+}5 \\ -77.847 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 8.2411e{+}5 \\ -214 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1.0924e{+}6 \\ -248.44 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1.1418e{+}6 \\ \hdashline 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \tag{134}$$

$$z_8 = \left(\begin{array}{cccc:cccc:c} -1 & 0 & 0 & 0 & 0.125 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0.125 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0.125 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0.125 & 0 \\ \hdashline \mathbf{-7.8374} & 1 & 0 & 0 & 0.29758 & 0 & 0 & 0 & \mathbf{3.0601e+5} \\ \mathbf{-77.706} & 0 & 1 & 0 & 0 & 0.99939 & 0 & 0 & \mathbf{8.209e+5} \\ \mathbf{-213.56} & 0 & 0 & 1 & 0 & 0 & 0.99953 & 0 & \mathbf{1.0878e+6} \\ \mathbf{-248.05} & 0 & 0 & 0 & 0 & 0 & 0 & 0.99977 & \mathbf{1.1398e+6} \\ \hdashline 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \tag{135}$$

$$z_9 = \left(\begin{array}{cccc:cccc:c} -1 & 0 & 0 & 0 & 0.125 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0.125 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0.125 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0.125 & 0 \\ \hdashline \mathbf{-0.038148} & 1 & 0 & 0 & 0.35114 & 0 & 0 & 0 & \mathbf{3.0601e+5} \\ \mathbf{-19.766} & 0 & 1 & 0 & 0 & 0.30858 & 0 & 0 & \mathbf{-1.6969e+6} \\ \mathbf{-81.871} & 0 & 0 & 1 & 0 & 0 & 0.66309 & 0 & \mathbf{1.0942e+6} \\ \mathbf{-245.99} & 0 & 0 & 0 & 0 & 0 & 0 & 0.99856 & \mathbf{1.1293e+6} \\ \hdashline 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \tag{136}$$

$$z_{10} = \left(\begin{array}{cccc:cccc:c} -1 & 0 & 0 & 0 & 0.125 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0.125 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0.125 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0.125 & 0 \\ \hdashline \mathbf{-12.857} & 1 & 0 & 0 & 0.93207 & 0 & 0 & 0 & \mathbf{3.0601e+5} \\ \mathbf{-76.947} & 0 & 1 & 0 & 0 & 0.99335 & 0 & 0 & \mathbf{8.0358e+5} \\ \mathbf{-212.92} & 0 & 0 & 1 & 0 & 0 & 0.99863 & 0 & \mathbf{1.081e+6} \\ \mathbf{-247.82} & 0 & 0 & 0 & 0 & 0 & 0 & 0.99963 & \mathbf{1.1386e+6} \\ \hdashline 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \tag{137}$$

$$z_{11} = \left(\begin{array}{cccc:cccc:c} -1 & 0 & 0 & 0 & 0.125 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0.125 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0.125 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0.125 & 0 \\ \hdashline \mathbf{-8.5941} & 1 & 0 & 0 & 0.99744 & 0 & 0 & 0 & \mathbf{3.0601e+5} \\ \mathbf{-35.284} & 0 & 1 & 0 & 0 & 0.41349 & 0 & 0 & \mathbf{-6.6087e+5} \\ \mathbf{-201.76} & 0 & 0 & 1 & 0 & 0 & 0.98646 & 0 & \mathbf{9.6616e+5} \\ \mathbf{-237.46} & 0 & 0 & 0 & 0 & 0 & 0 & 0.99346 & \mathbf{1.0869e+6} \\ \hdashline 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \tag{138}$$

$$\left.\frac{\delta \bar{H}}{\delta Z}\right|_{Z_{11}} = \left(\begin{array}{cccc:cccc:c} 0.0070996 & 0.42715 & 0.70617 & 0.23434 & 0.056797 & 3.4172 & 5.6494 & 1.8748 & 1.5951e{-6} \\ 0.026111 & 1.5582 & 2.559 & 0.78988 & 0.20889 & 12.465 & 20.472 & 6.3191 & 5.8253e{-6} \\ 0.0055611 & 0.33185 & 0.545 & 0.16821 & 0.044489 & 2.6548 & 4.36 & 1.3456 & 1.2406e{-6} \\ 0.021527 & 1.2755 & 2.0826 & 0.59689 & 0.17222 & 10.204 & 16.661 & 4.7751 & 4.7731e{-6} \\ \hdashline \mathbf{0.026111} & 1.5582 & 2.559 & 0.78988 & 0.20889 & 12.465 & 20.472 & 6.3191 & \mathbf{5.8253e{-6}} \\ \mathbf{0.0055611} & 0.33185 & 0.545 & 0.16821 & 0.044489 & 2.6548 & 4.36 & 1.3456 & \mathbf{1.2406e{-6}} \\ \mathbf{0.021527} & 1.2755 & 2.0826 & 0.59689 & 0.17222 & 10.204 & 16.661 & 4.7751 & \mathbf{4.7731e{-6}} \\ \mathbf{0.12202} & 7.1531 & 11.576 & 2.8809 & 0.97617 & 57.225 & 92.611 & 23.047 & \mathbf{2.6807e{-5}} \\ \hdashline 30.319 & 1777.3 & 2876.4 & 715.84 & 242.55 & 14219 & 23011 & 5726.7 & 0.0066609 \end{array}\right) \tag{139}$$

$$\left.\frac{\delta\,|\bar{\lambda}|}{\delta Z}\right|_{Z_{11}} = \left(\begin{array}{cccc:cccc:c} 0.41162 & 2.5905 & 14.134 & 20.308 & 3.2929 & 20.724 & 113.08 & 162.46 & 5.752e{-6} \\ 0.11223 & 4.3006 & 6.5898 & 4.4816 & 0.8978 & 34.405 & 52.718 & 35.853 & 1.5958e{-5} \\ 0.036791 & 0.96679 & 1.4889 & 0.81767 & 0.29433 & 7.7343 & 11.911 & 6.5413 & 3.5981e{-6} \\ 0.016307 & 0.91831 & 0.46321 & 1.2742 & 0.13046 & 7.3465 & 3.7057 & 10.194 & 3.3907e{-6} \\ \hdashline \mathbf{0.11223} & 4.3006 & 6.5898 & 4.4816 & 0.8978 & 34.405 & 52.718 & 35.853 & \mathbf{1.5958e{-5}} \\ \mathbf{0.036791} & 0.96679 & 1.4889 & 0.81767 & 0.29433 & 7.7343 & 11.911 & 6.5413 & \mathbf{3.5981e{-6}} \\ \mathbf{0.016307} & 0.91831 & 0.46321 & 1.2742 & 0.13046 & 7.3465 & 3.7057 & 10.194 & \mathbf{3.3907e{-6}} \\ \mathbf{0.1568} & 9.4818 & 15.968 & 3.9284 & 1.2544 & 75.855 & 127.75 & 31.427 & \mathbf{3.5425e{-5}} \\ \hdashline 39.27 & 2367.2 & 3785.7 & 679.78 & 314.16 & 18938 & 30286 & 5438.3 & 0.0088362 \end{array}\right) \tag{140}$$

$$A_p = \begin{pmatrix} 8.384e{-1} & 1.600e{-1} & -3.294e{-1} & -4.833e{-2} & 0 & 0 & 0 & 0 & 0 & 0 \\ -3.927e{-1} & 7.144e{-1} & 5.040e{-2} & -8.245e{-3} & 0 & 0 & 0 & 0 & 0 & 0 \\ -1.566e{-1} & -6.105e{-1} & 3.683e{-2} & 4.195e{-1} & 0 & 0 & 0 & 0 & 0 & 0 \\ -1.444e{-1} & 1.772e{-1} & -6.798e{-1} & 6.508e{-1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 1.929e{-1} & 1.512e{-1} & 4.030e{-1} & 3.898e{-1} & 9.773e{-1} & 1.037e{-2} & -6.170e{-2} & 0 & 0 & 0 \\ 2.768e{-4} & 2.170e{-4} & 5.783e{-4} & 5.594e{-4} & 2.837e{-3} & 9.971e{-1} & 1.698e{-2} & 0 & 0 & 0 \\ 3.238e{-2} & 2.539e{-2} & 6.767e{-2} & 6.545e{-2} & 3.320e{-1} & -3.341e{-1} & 9.868e{-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.000e{+0} & -1.000e{-10} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.000e{-2} & 1.000e{+0} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 9.417e{-1} \end{pmatrix} \tag{141}$$

$$B_p = \begin{pmatrix} -4.007e{+0} \\ -5.769e{+0} \\ -6.522e{+0} \\ 2.490e{+0} \\ 8.562e{-1} \\ 1.229e{-3} \\ 1.438e{-1} \\ 1.000e{+0} \\ 5.000e{-3} \\ 0 \end{pmatrix}, \quad C_p = \begin{pmatrix} 9.209e{-3} \\ 7.221e{-3} \\ 1.924e{-2} \\ 1.861e{-2} \\ 9.441e{-2} \\ 4.953e{-4} \\ -2.946e{-3} \\ 0 \\ 0 \\ -3.495e{-1} \end{pmatrix}^{\top}, \quad B = \begin{pmatrix} -2.372e{+0} \\ -2.540e{+0} \\ -1.210e{-1} \\ -1.565e{-4} \\ -6.245e{-2} \\ 1.151e{+0} \\ 4.083e{-2} \\ 2.255e{-1} \\ -1.528e{-2} \\ -9.720e{-4} \end{pmatrix}, \quad C = \begin{pmatrix} -2.372e{-2} \\ 2.540e{-2} \\ 1.210e{-3} \\ -1.565e{-6} \\ 6.245e{-4} \\ 1.151e{-2} \\ 4.083e{-4} \\ -2.255e{-3} \\ 1.528e{-4} \\ 9.720e{-6} \end{pmatrix}^{\top}, \quad D = -2.140e{-1} \tag{142}$$

$$A = \begin{pmatrix} 8.195e{-1} & 2.812e{-1} & -3.317e{-2} & 2.699e{-2} & -1.649e{-1} & 1.318e{-1} & 1.059e{-2} & -6.733e{-2} & 1.750e{-3} & 6.525e{-5} \\ -2.812e{-1} & -4.817e{-1} & -1.668e{-1} & 8.654e{-2} & -5.403e{-1} & 1.469e{-1} & 1.837e{-2} & -1.211e{-1} & 1.942e{-3} & 2.134e{-5} \\ 3.317e{-2} & -1.668e{-1} & 9.749e{-1} & 1.696e{-2} & -9.104e{-2} & 7.638e{-2} & 3.357e{-3} & -2.006e{-2} & 8.441e{-4} & 4.548e{-5} \\ 2.699e{-2} & -8.654e{-2} & -1.696e{-2} & 9.601e{-1} & 2.528e{-1} & 5.956e{-2} & 1.654e{-3} & -9.085e{-3} & 6.046e{-4} & 3.843e{-5} \\ 1.649e{-1} & -5.403e{-1} & -9.104e{-2} & -2.528e{-1} & 6.022e{-1} & 3.888e{-1} & 1.150e{-2} & -6.420e{-2} & 3.945e{-3} & 2.454e{-4} \\ 1.318e{-1} & -1.469e{-1} & -7.638e{-2} & 5.956e{-2} & -3.888e{-1} & 4.664e{-1} & -6.206e{-2} & 4.224e{-1} & -8.490e{-4} & 3.703e{-4} \\ 1.059e{-2} & -1.837e{-2} & -3.357e{-3} & 1.654e{-3} & -1.150e{-2} & -6.206e{-2} & 9.832e{-1} & 1.258e{-1} & 7.737e{-3} & 6.392e{-4} \\ 6.733e{-2} & -1.211e{-1} & -2.006e{-2} & 9.085e{-3} & -6.420e{-2} & -4.224e{-1} & -1.258e{-1} & -4.483e{-2} & 7.258e{-2} & 5.631e{-3} \\ -1.750e{-3} & 1.942e{-3} & 8.441e{-4} & -6.046e{-4} & 3.945e{-3} & 8.490e{-4} & -7.737e{-3} & 7.258e{-2} & 9.838e{-1} & -2.474e{-3} \\ -6.525e{-5} & 2.134e{-5} & 4.548e{-5} & -3.843e{-5} & 2.454e{-4} & -3.703e{-4} & -6.392e{-4} & 5.631e{-3} & -2.474e{-3} & 9.418e{-1} \end{pmatrix} \tag{143}$$